ITERATIVE METHODS FOR ELLIPTIC DIFFERENCE EQUATIONS

R. P. Fedorenko

Contents

§ 1. Introduction

The method of finite differences is proving to be the most powerful and universal tool of the approximate solution of complicated boundary-value problems in elliptic linear equations. This is because its flexibility is sufficient to cope with problems that cause considerable difficulties in classical approximation methods.[1] True, it requires considerably more simple and uniform arithmetic operations, but this difficulty is overcome by the increased speed of electronic computers and the improvement of algorithms. It should be said that the progress made in electronic computers does not diminish the value of the most effective algorithms. As the increased capacity of the memory permits the solution of more complicated problems, the amount of work would grow, without the use of the most efficient algorithms, significantly faster than the speed of computers. There are two separate questions which occur in solving an elliptic equation by the finite-difference method.

[1] Such as the Ritz, Galerkin, Fourier, potential theory methods, etc.

1. Construction of mesh spaces, the approximation of the differential operator[1] by various difference operators (the choice of the difference scheme), and a justification of the method (a proof that the finite-difference solutions tend to the solution of the original equation). We completely ignore this side of the problem in the present survey, though of course we use some (but only the simplest and most obvious) difference schemes in what follows.

2. The actual solution of the finite-difference equation. Formally, we are concerned with a system of linear algebraic equations of a specific shape (each row of the matrix has only a few non-zero entries whose position is given by a simple rule) and a very high order (say $10^3 - 10^4$).

Some very efficient methods have been developed for the solution of these special systems (they could be called elliptic difference equations). The present paper surveys the main achievements in this area.

A few preliminary remarks on the character of the exposition.

The progress in this area of numerical analysis has been determined by several fundamental 'inventions'. They are, as a matter of fact, not many, and it is on them that we concentrate our attention, explaining them on the simplest concrete material, such as Poisson's equation in the square. Every such 'invention' naturally creates an endeavour to use the underlying idea in the largest possible class of problems, perhaps with some modifications of the algorithm. These generalizations are also of great interest, and we try to give a sufficiently clear picture of the achievements in that area, as well as of the difficulties that have not yet been overcome. Thus, we are mainly concerned with the simplest example: Poisson's equation

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=f(x, y) \tag{1.1}$$

in the domain

$$(0 \leqslant x < \pi) \times (0 \leqslant y \leqslant \pi)$$

with boundary conditions of the first kind:

$$u|_\Gamma = \varphi(s). \tag{1.2}$$

However, we are also interested to what extent and with what effect one method or another is applicable in the following more complicated problems:

1. Transition to boundary conditions, for example, of the third kind:

$$\alpha \frac{\partial u}{\partial n}+\beta u\Big|_\Gamma = \varphi(s). \tag{1.3}$$

2. Transition to equations with three (or more) independent variables.

3. Transition to boundary-value problems in more complicated domains, for example, a domain bounded by a piecewise smooth curve (or surface).

[1] Including boundary conditions.

4. Transition to a more general differential equation with variable coefficients, mixed and lower derivatives:

$$\frac{\partial}{\partial x} a(x, y) \frac{\partial u}{\partial x}+\frac{\partial}{\partial y} b(x, y) \frac{\partial u}{\partial x}+\frac{\partial}{\partial x} b(x, y) \frac{\partial u}{\partial y}+\frac{\partial}{\partial y} c(x, y)+ +A \frac{\partial u}{\partial x}+B \frac{\partial u}{\partial y}+Cu=f. \tag{1.4}$$

In this survey we try to concentrate on those generalizations of the fundamental constructions that lead to a significant progress in the directions listed above. We do not consider the generalizations concerning directional derivatives, boundary-value problems for higher order equations (for example, biharmonic), or systems of elliptic equations, although the main ideas of the methods described are also applicable to them.

The paper contains comparatively few proofs, which are selected with the following points in mind: 1. They reflect the main stages in the development of the iterative methods. 2. They are elementary from the point of view of the tools used. As for various generalizations requiring a more refined technique, we essentially limit ourselves to an exposition of the results.

Great attention is paid to the meaningful evaluation of results (from the point of view of a computer programmer, who is interested in the effectiveness of the method and in its applicability to various concrete problems). We wish to give the reader a general idea of the state of affairs in this part of numerical analysis without averting his attention to the technicalities. Naturally, we give references to papers containing the proofs in full.

This survey is primarily addressed to the specialist who is faced with the necessity of solving elliptic problems and who would like to know the resources of numerical mathematics. The author hopes that the survey will help such a reader to choose a suitable method, perhaps modifying it to fit the application to his problem. The notation used in the paper is also directed towards such a reader. The author has striven to achieve the maximal mnemonic expressiveness. However, it should be borne in mind that the language in current journals is different. The scope of new constructions is not described in terms of concrete problems (the type of equations, domain, boundary conditions), but in abstract terms of the theory of linear operators (positiveness, decomposition into a sum of positive operators, into a sum of commuting operators etc.).

To facilitate the reading of journals, we supplement the exposition of various methods by the corresponding abstract formulations.

The author hopes that this survey will also prove useful to those for whom computational algorithms are primarily the object of theoretical research. The author's aim is to show in what way the results of this research are valuable in practice and where the need to remove the difficulties is now greatest.

§ 2. Statement of the problem; basic notation

Let us consider the finite-difference method of solving the first boundary-value problem for the equation $\Delta u = f$ in the $\pi \times \pi$ square ((1.1), (1.2)). We introduce a rectangular mesh $\{x_n = nh, y_m = mh, n, m = 0, 1, \ldots, N\}$ (for simplicity, we consider a uniform mesh with the same number of points in x and y). For the approximate solution we take the function $u_{n,m}$ $(n, m = 0, 1, \ldots, N)$ defined on the vertices of the mesh and satisfying the difference equations

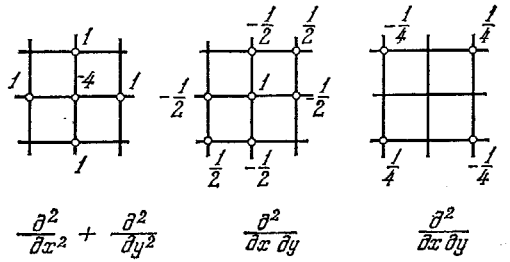


Fig. 1.

$$(2.1)\qquad \left(\frac{\partial^2 u}{\partial x^2}\right)_{n,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)_{n,m} = f_{n,m} \quad (n,\ m = 1,\ 2,\ \ldots,\ N-1),$$

together with boundary conditions (of the first kind)

$$(2.2)\qquad \begin{cases} u_{n,0} = \varphi_{n,0},\ u_{n,N} = \varphi_{n,1} \\ \quad (n = 0,\ 1,\ \ldots,\ N), \\ u_{0,m} = \varphi_{0,m},\ u_{N,m} = \varphi_{1,m} \\ \quad (m = 0,\ 1,\ \ldots,\ N) \end{cases}$$

(where f and φ are given mesh functions defined, respectively, on the interior and boundary vertices). Let us clarify the notation: unless the contrary is explicitly stated, we use in (2.1) the simplest difference approximation, that is,

$$(2.3)\qquad \left(\frac{\partial^2 u}{\partial x^2}\right)_{n,m} \equiv \frac{u_{n-1,m} - 2u_{n,m} + u_{n+1,m}}{h^2}.$$

If the equation has variable coefficients, then

$$(2.4)\qquad \left(\frac{\partial}{\partial x} a \frac{\partial u}{\partial x}\right)_{n,m} \equiv \equiv \frac{1}{h}\left[a_{n+\frac{1}{2},m}\frac{u_{n+1,m} - u_{n,m}}{h} - a_{n-\frac{1}{2},m}\frac{u_{n,m} - u_{n-1,m}}{h}\right].$$

For the mixed derivative we use, for example, the scheme

$$(2.5)\qquad \left(\frac{\partial^2 u}{\partial x\,\partial y}\right)_{n,m} \equiv \frac{1}{2h}\left[\frac{u_{n+1,m+1} - u_{n+1,m-1}}{2h} - \frac{u_{n-1,m+1} - u_{n-1,m-1}}{2h}\right],$$

or

$$(2.6)\qquad \left(\frac{\partial^2 u}{\partial x\,\partial y}\right)_{n,m} \equiv \frac{1}{2h}\left[\frac{u_{n+1,m+1} - u_{n+1,m}}{h} - \frac{u_{n,m+1} - u_{n,m}}{h}\right] + \frac{1}{2h}\left[\frac{u_{n,m} - u_{n,m-1}}{h} - \frac{u_{n-1,m} - u_{n-1,m-1}}{h}\right].$$

The stencils of the difference schemes (2.1), (2.5) and (2.6) are given in Fig. 1, which shows the vertices used in approximating the differential expression in the vertex (n, m), and the corresponding weights (multiplied by h^2).

We sometimes write (2.1) in the form

$$(2.1^*)\qquad (\Delta u)_{n,m} = f_{n,m}.$$

With these explanations the reader should have no difficulty in finding the difference operator

$$(2.7)\qquad \left(\frac{\partial^2}{\partial x^2}\frac{\partial^2 u}{\partial y^2}\right)_{n,m}$$

in the simplest nine-vertex scheme.

For the system (2.1) – (2.2) it is easy to prove the existence and uniqueness of the solution and its continuous dependence on f and φ (with estimates that are uniform in h, which is very important for the convergence of numerical solutions to the exact solution). Below we describe various methods for solving the system (2.1)–(2.2) with $\sim N^2$ unknowns $u_{n,m}$. These are, essentially, iterative methods, where we start with some initial approximation $u^0_{n,m}$ and successively obtain the values $u^1_{n,m}, u^2_{n,m}, \ldots, u^\nu_{n,m}, \ldots$, so that

$$\lim_{\nu\to\infty} u^\nu_{n,m} = u^*_{n,m}$$

(where $u^*_{n,m}$ is the exact solution of the difference equations (2.1)–(2.2)). However, we ought here to make a reservation as to the use of the term 'exact solution of the difference problem'. The reason is that all the computations are performed on a machine with finitely many decimal places (usually between 8 and 12). Let $U(x, y)$ be the exact solution of the original problem (1.1)–(1.2), $U_{n,m} = U(x_n, y_m)$ and $U^*_{n,m}$ the computer representation of $U_{n,m}$, that is, $U^*_{n,m} = U_{n,m}(1 + \varepsilon_{n,m})$, where $\varepsilon_{n,m}$ is the rounding error, which is a random variable with $|\varepsilon| \sim 10^{-8}$–$10^{-12}$, depending on the length of the memory cell.

We substitute $U^*_{n,m}$ in the difference equation (2.1):

$$(2.8)\qquad (\Delta U^*)_{n,m} - f_{n,m} \simeq Ah^2 + B\frac{\varepsilon}{h^2} + \varepsilon F.$$

Here Ah^2 is the approximation error, $A \sim \left|\frac{\partial^4 U}{\partial x^4}\right|, \left|\frac{\partial^4 U}{\partial y^4}\right|$ (assuming that U is continuously differentiable sufficiently many times), and B and F are quantities depending on the functions U and f. The formula (2.8) makes it possible to arrive at certain qualitative conclusions, which are important in computational practice.

First of all, it is clear that h cannot be diminished without increasing the number of decimal places: a reasonable lower bound for h is determined by requiring that the approximation error Ah^2 and the rounding error

$B\frac{\varepsilon}{h^2}$ are of the same order, that is, $h \simeq \left[\frac{B\varepsilon}{A}\right]^{1/4}$. (If we use fourth order difference approximation, that is, one with the approximation error $O(h^4)$, we obtain in the same way the lower bound $h \geqslant O(\varepsilon^{1/6})$ for the size of the mesh step.)

Secondly, it is obvious that (2.1) has no exact solution in the class of machine numbers, and that in fact there is no need for it. Any approximate solution is characterized by the size of the 'residual'

$$r^{\nu}_{n,m} \equiv (\Delta u^{\nu})_{n,m} - f_{n,m} \quad (n, m = 1, 2, \ldots, N-1). \tag{2.9}$$

There is no need to continue with the iterations after an approximate solution u^{ν} with $\|r^{\nu}\| \sim Ah^2$ is obtained. We are mainly interested in the convergence of the iterations for $\|r^{\nu}\| \gg Ah^2 > \frac{B\varepsilon}{h^2}$, and so the rounding errors are ignored in the first stage of the study of an iterative process. This may lead to the wrong conclusions, in which case a more accurate analysis has to be performed. Thus, we judge upon the accuracy of the approximation from the residual norm[1]

$$\|r^{\nu}\| \equiv \left\{ \sum_{n=1}^{N-1} \sum_{m=1}^{N-1} h^2 (r^{\nu}_{n,m})^2 \right\}^{1/2}. \tag{2.10}$$

In this way we introduce the usual inner product and norm in the space of mesh functions. The accuracy of the approximate solution can also be measured by the norm of the error

$$v^{\nu}_{n,m} \equiv u^{\nu}_{n,m} - u^{*}_{n,m} \tag{2.11}$$

(where $u^{*}_{n,m}$ is the exact solution of (2.1)–(2.2); we recall that so far we ignore the rounding errors and assume that $u^{*}_{n,m}$ exists). To characterize the accuracy of u^{ν} by the quantity $\|v^{\nu}\|$ is convenient in experimental work, when $u^{*}_{n,m}$ is known (for example, we may start with an arbitrary function $u^{*}_{n,m}$, determine f and φ from (2.1) and use them in the computation). In practical work, however, the achieved accuracy is usually measured by $\|r^{\nu}\| = \|\Delta u^{\nu} - f\|$. It is useful to note that v^{ν} satisfies the homogeneous boundary conditions and

$$(\Delta v^{\nu})_{n,m} = (\Delta u^{\nu})_{n,m} - f_{n,m} = r^{\nu}_{n,m} \quad (n, m = 1, 2, \ldots, N-1). \tag{2.12}$$

The analysis of the convergence of iteration processes depends very much on the spectra analysis of certain operators. We introduce the eigenfunctions and the eigenvalues of the matrix of the difference equation

[1] The accuracy of u^{ν} is sometimes measured by $\max_{m,n} \frac{|u^{\nu+1}_{m,n} - u^{\nu}_{m,n}|}{\|u^{\nu}\|}$, that is, by the stabilization of u^{ν}. One should remember, however, that a good stabllization of u^{ν} may be due to very slow convergence.

(2.1), together with the homogeneous boundary conditions

$$\begin{cases} (\Delta\varphi^{(p)})_{n,m} = -\lambda_p \varphi^{(p)}_{n,m}, & \|\varphi^{(p)}\| = 1, \\ \varphi^{(p)}_{n,m} = 0 \text{ on the boundary.} \end{cases} \tag{2.13}$$

In (2.13) we often use separation of variables, in which case it is convenient to label the eigenfunctions and eigenvalues by pairs of indices

$$\Delta\varphi^{(p,q)} = -\lambda_{p,q}\varphi^{(p,q)}.$$

Here

$$\lambda_{p,q} = \lambda'_p + \lambda''_q, \quad \varphi^{(p,q)}_{n,m} = \psi^{(p)}_n \chi^{(q)}_m$$

and

$$\left(\frac{\partial^2 \psi^{(p)}}{\partial x^2}\right)_n = -\lambda'_p \psi^{(p)}_n, \quad \left(\frac{\partial^2 \chi^{(q)}}{\partial y^2}\right)_m = -\lambda''_q \chi^{(q)}_m.$$

For the simplest problem (2.1)–(2.2) the exact values of λ'_p, λ''_q, $\psi^{(p)}_n$ and $x^{(q)}_m$ are well known,[1] but we do not make any use of them. For a later characterization of a spectrum its bounds are essential:

$$0 < l \leqslant \lambda_{p,q} \leqslant L, \quad 0 < l' \leqslant \lambda'_p \leqslant L', \quad 0 < l'' \leqslant \lambda''_q \leqslant L''. \tag{2.14}$$

The lower bounds l, l' and l'' coincide up to $O(h^2)$ with the lower spectral bounds of the corresponding differential operators; the upper bounds L', L'' and L are usually well estimated by such quantities as the maximum (in m and n) of the sum of the moduli of the weight coefficients in the difference scheme. For (2.1) this gives the nearly exact value $L \simeq \frac{8}{h^2} = \frac{8N^2}{\pi^2}$. An efficient application of iterative methods to the solution of finite-difference elliptic equations requires a reasonably accurate estimation of l (below) and L (above). Estimates for l may be obtained either by one of the classical methods or by essentially the same iterative method as that used in solving the problem. We have now introduced some of the fundamental concepts that are used in what follows; the remainder appear as we describe the iterative methods.

§ 3. The simplest iterative methods

Taken on their own, the methods in this section are relatively unimportant, due to their very slow convergence (for values of $N \sim 50 - 100$, typical in present-day computations). They are, however, used as elements of more perfect algorithms; furthermore, their exposition is a simple and convenient way of introducing some of the concepts that are important in what follows.

[1] For example, $\psi^{(p)}_n = \sin\frac{np\pi}{N}$ (ignoring the normalizing factor).

1. The method of simple iteration. We have

$$(3.1)\quad \begin{cases} u^{\nu+1}_{n,m} = u^{\nu}_{n,m} + \tau\,[(\Delta u^{\nu})_{n,m} - f_{n,m}], & n,\ m = 1,\ 2,\ \ldots,\ N-1, \\ u^{\nu+1}_{n,m} = u^{\nu}_{n,m} \text{ on the boundary} \end{cases}$$

(assuming that u^0 satisfies the boundary conditions). Henceforth we use a shortened expression for (3.1)

$$(3.1^*)\qquad u^{\nu+1} = u^{\nu} + \tau(\Delta u^{\nu} - f).$$

Subtracting the obvious equation $u^* = u^* + \tau(\Delta u^* - f)$ from (3.1*) we obtain a formula for the evolution of the error v^ν in the iterative process

$$(3.2)\quad \begin{cases} v^{\nu+1} = v^{\nu} + \tau\cdot\Delta v^{\nu} = (E + \tau\Delta)\,v^{\nu}, \\ v^{\nu+1} = 0 \text{ on the boundary.} \end{cases}$$

We expand v^0 as a Fourier series in the eigenfunctions of Δ:

$$v^0 = \sum_p c_p \varphi^{(p)}.$$

Then, obviously,

$$(3.3)\qquad v^{\nu} = \sum_p c_p (1 - \tau\lambda_p)^{\nu} \varphi^{(p)}.$$

Furthermore, $\|v^0\| = [\sum_p c_p^2]^{1/2}$ and

$$(3.4)\qquad \|v^{\nu}\| = [\sum_p c_p^2 (1-\tau\lambda_p)^{2\nu}]^{1/2} \leqslant \|v^0\| \max_{l\leqslant\lambda\leqslant L} |1-\tau\lambda|^{\nu}.$$

From (3.4) it is clear that for the greatest efficiency of simple iterations the parameter τ must be a solution of the problem

$$(3.5)\qquad \min_{\tau} \max_{l\leqslant\lambda\leqslant L} |1-\tau\cdot\lambda|.$$

The optimal value τ^* is easily found:

$$(3.6)\qquad \tau^* = \frac{2}{L+l}$$

and gives the coefficient of convergence

$$\mu = \max_{l\leqslant\lambda\leqslant L} |1-\tau^*\lambda| = 1-\tau^* l = -(1-\tau^* L) = \frac{L-l}{L+l} \simeq 1 - 2\frac{l}{L} \simeq e^{-2\frac{l}{L}}.$$

The quantity $\varkappa = -\ln \mu$ is called the index of the quality of the iterative process, since it determines how many iterations are needed to obtain an approximation u^ν with the norm of the error ε^{-1} times smaller than that of the initial approximation u^0. For it follows from $\|v^\nu\| \leqslant \varepsilon\|v^0\|$ and (3.4) that

$$(3.7)\qquad \nu \simeq \frac{\ln \varepsilon^{-1}}{\varkappa} \simeq \frac{L}{2l}\ln\frac{1}{\varepsilon}.$$

This fact is sometimes expressed by a formula for the decrease of $\|v^\nu\|$ in the process of iterations:

$$(3.8)\qquad \|v^{\nu}\| \simeq \|v^0\| e^{-\varkappa\cdot\nu}.$$

Applying the difference operator Δ to both sides of (3.2) or (3.3) we see that the residual norm decreases at the same rate

$$(3.9)\qquad \|r^{\nu}\| \leqslant \mu^{\nu}\|r^0\|, \quad \text{i.e.} \quad \|r^{\nu}\| \simeq \|r^0\| e^{-\varkappa\cdot\nu}.$$

For the problem (2.1)–(2.2) we have $l = 2 + O(h^2)$, $L \simeq 8N^2/\pi^2$, and $\varkappa = \pi^2/2N^2$. Taking, for example, $N = 100$, and assuming that it is necessary to diminish the original residuals 10^5 times, we find the number of iterations to be $\nu \simeq 2{,}5\cdot 10^4$. A computer with $\sim 10^5$ operations per second performs one iteration (computation of 10^4 values in the vertices of the mesh) in about 0.5 sec., and the complete calculation needs more than 3 hours of machine time, which is totally unacceptable. Of course, the real evolution of $\|r^\nu\|$ is not quite the same as in (3.9). The iterations $\|r^\nu\|$ decrease at first at a much faster rate, due to the decrease in those terms of (3.3) for which $|1 - \tau\lambda_p| \simeq 0$. This can be described as an intensive suppression of those components of the error that correspond to the middle part of the spectrum. However, relatively soon (3.3) effectively contains only those terms that correspond to the points of the spectrum close to the bounds l and L, and thereafter $\|r^\nu\|$ decreases according to (3.8). The observation that the simple iterations converge extremely slowly at the boundary of the spectrum and more quickly in its middle is the starting point for Richardson's method of improving the rate of convergence. Note also that for $\tau \simeq 1/L$ the rate of convergence depends asymptotically only on the lower bound of the spectrum:

$$\|r^{\nu}\| \simeq \|r^0\| \left(1 - \frac{l}{L}\right)^{\nu} = \|r^0\|\,(1-\tau l)^{\nu}.$$

This permits the use of iterations with $\tau \simeq 1/L$ in order to determine the approximate value of l: we choose integers ν_1 and ν_2 and, after $\nu_1 + \nu_2$ iterations, put

$$(3.10)\qquad l \simeq \frac{1}{\tau}\left[1 - \left(\frac{\|r^{\nu_1+\nu_2}\|}{\|r^{\nu_1}\|}\right)^{1/\nu_2}\right].$$

This formula becomes more accurate with growing ν_1, and it would be precise if (3.3) contained only one summand, corresponding to $\lambda_p = l$. The presence of other summands means that $\|r^\nu\|$ decreases at a faster rate than $e^{-\nu\tau l}$, so that (3.10) gives an excessive value for l. The problem of determining l is facilitated by the use of meshes with a small number of vertices, much fewer than N^2. This is due to the fact that l is an approximation of the first eigenvalue of the original differential problem and that the first eigenfunction is the smoothest and gives the closest fit between the differential

and difference operators.

All this can be stated as the following theorem.

Theorem 1 *The method of simple iteration:*

$$u^{\nu+1} = u^{\nu} + \tau(\Delta u^{\nu} - f)$$

for the problem (2.1)–(2.2) *with* $\tau = \frac{2}{L+l}$ *converges for any initial function* u^0. *The residual norm and the error decrease in the iteration process according to the formulae*

$$\|v^{\nu}\| \simeq \|v^0\| e^{-2\nu l/L}, \qquad \|r^{\nu}\| \simeq \|r^0\| e^{-2\nu l/L}.$$

In order to decrease the residual of the original approximation ε^{-1} *times, it is necessary to perform* $\nu \simeq \frac{L}{2l} \ln \frac{1}{\varepsilon}$ *iterations.*

For the difference approximations of the second order elliptic equations we usually have $\frac{L}{l} \simeq O(N^2)$, and so the number of iterations[1] is $O\left(N^2 \ln \frac{1}{\varepsilon}\right)$.

The estimates of the decrease of $\|v^{\nu}\|$ and $\|r^{\nu}\|$ cannot be improved if we admit arbitrary initial functions u^0; for it is sufficient to take $u^0 = u^* + c\varphi$, where φ is the eigenfunction corresponding to $\lambda = l$ (or L) (here and in what follows we assume that $l \ll L$ and we disregard the difference between $1 - 2l/L$ and $e^{-2l/L}$).

It is easy to see that we have not in fact used the concrete form of the difference operator Δ. We only need two of its properties:

1. Self-adjointness (as we have used the machinery of the Fourier series).
2. Positiveness of the difference operator $-\Delta$; to perform the computations and estimate their efficiency one need only know the spectral bounds.

Hence the method of simple iteration withstands the widest range of generalizations of the problem in all the directions we are interested in. It is, apparently, the most general method for solving elliptic difference problems. Here is another problem concerning the construction of the approximating difference operators: suppose that the original problem has the form

$$Du = f, \quad \Gamma u = \varphi.$$

We do not specify the form of the equation, the form of the boundary conditions, or the shape of the domain. We only assume that the operator $-D$ with boundary conditions $\Gamma u = 0$ is linear, positive, and self-adjoint. To what extent are these properties preserved under one finite-difference approximation or another? There exist devices that enable us to obtain

[1] The number of iterations is sometimes given as $O\left(\frac{1}{h^2} \cdot \ln \frac{1}{\varepsilon}\right)$. This is not a very good formulation, because what matters here is the number of vertices and not the size of the step.

difference operators necessarily having these properties.[1] On the other hand, it is known that certain difference approximations may destroy self-adjointness; at the same time it is clear that any such loss must be of the order of the approximation error. Bearing in mind this, and the unavoidable use of methods in practical computations beyond the theoretically studied situations, we need not be too concerned about the lack of self-adjointness of this kind. All the more, because in the convergence proof there is one so far unused resource: for the convergence of the iterations

$$u^{\nu+1} = u^{\nu} + \tau(Du^{\nu} - f)$$

it is sufficient to have an estimate of the norm of the iteration operator

$$||E + \tau D|| = \mu < 1.$$

A more serious difficulty arises in the generalization of the problem for which the spectrum of the operator crosses the negative semi-axis, that is, $l < 0$. This may be due to lower order terms in the initial differential equation. Such problems are of interest and we shall return to them.

It is appropriate here to give a short explanation of the boundary conditions of, say, the second or third kind. To be specific, consider the equation (2.1) with the boundary conditions $\frac{\partial u}{\partial x} = \varphi$ for $x = 0$, or, in finite differences,

$$\frac{u_{1,m} - u_{0,m}}{h} = \varphi_m \qquad (m = 1, 2, \ldots, N-1).$$

There are two ways of realizing such a boundary condition in the iteration process.

1. Having computed $u^{\nu+1}_{n,m} = u^{\nu}_{n,m} + \tau(\Delta u^{\nu} - f)_{n,m}$ at all interior points $(n, m = 1, 2, \ldots, N-1)$, find $u_{0,m}$ $(m = 1, \ldots, N-1)$ from the formula

$$u_{0,m} = u_{1,m} - h\varphi_m.$$

2. First take the boundary conditions into account, transforming the difference equations at the points $(1, m)$ in an obvious way. Then the iteration formula for the quantities $u^{\nu}_{1,m}$ takes the form

$$u^{\nu+1}_{1,m} = u^{\nu}_{1,m} + \tau\left[\frac{u^{\nu}_{2,m} - u^{\nu}_{1,m}}{h^2} - \frac{\varphi_m}{h} + \left(\frac{\partial^2 u^{\nu}}{\partial y^2}\right)_{1,m}\right].$$

The author has used both methods in his computations and, although he has a distinct preference for the second, cannot raise any serious objections against the first.

II Seidel's method. This method, which is close in efficiency to that of the simple iteration, is conveniently introduced as the simplest version of the so-called method of minimal residuals. Its idea is simple and sufficiently

[1] These are the so-called variation-difference schemes; however, they are not always convenient in computations.

fruitful, since in combination with certain non-trivial constructions, it leads to efficient methods of solving finite-difference elliptic problems. It is well known that solutions of many boundary-value problems for elliptic differential equations minimize a certain functional. For example, the solution of the first boundary-value problem for the equation

$$(3.11)\qquad Du \equiv \frac{\partial}{\partial x} a \frac{\partial u}{\partial x} + \frac{\partial}{\partial y} b \frac{\partial u}{\partial x} + \frac{\partial}{\partial x} b \frac{\partial u}{\partial y} + \frac{\partial}{\partial y} c \frac{\partial u}{\partial y} = f, \quad u|_\Gamma = \varphi,$$

minimizes (in the class of functions satisfying the boundary condition $u|\Gamma = \varphi$) the value of the functional (Dirichlet integral)

$$(3.12)\qquad F[u(\cdot,\cdot)] \equiv \frac{1}{2}\iint [au_x^2 + 2bu_x u_y + cu_y^2]\,dx\,dy + \iint fu\,dx\,dy.$$

Computing in the obvious way the functional derivative of F with respect to $u(.,.)$ and equating it to zero, we obtain (3.11). In other words,

$$(3.13)\qquad \frac{\partial F[u(\cdot,\cdot)]}{\partial u(\cdot,\cdot)} = -Du + f.$$

Another functional with a self-adjoint operator D can also be used:

$$(3.14)\qquad F[u(\cdot,\cdot)] \equiv \frac{1}{2}\iint (-Du + 2f)\,u\,dx\,dy.$$

For this we also have (3.13). Using the space of the mesh-functions $u_{n,m}$ and some quadratic formula for the functional F (3.12), we can compute the derivative F_u and obtain the finite-difference analogue of the equation (3.13) with a self-adjoint difference operator that approximates the differential operator D. This is the main idea in constructing variational difference schemes. Thus, the simplest five-point difference operator for the Laplace equation is obtained by taking for $F[u]$ the expression

$$(3.15)\qquad F[u] \equiv \frac{1}{2}h^2 \sum_{m=1}^{N-1}\sum_{n=0}^{N-1} \left(\frac{u_{n+1,m} - u_{n,m}}{h}\right)^2 + \frac{1}{2}h^2 \sum_{m=0}^{N-1}\sum_{n=1}^{N-1} \left(\frac{u_{n,m+1} - u_{n,m}}{h}\right)^2 + \sum_{n=1}^{N-1}\sum_{m=1}^{N-1} h^2 f_{n,m} u_{n,m}.$$

In the method of minimal residuals, having obtained some approximate solution u^ν we choose a function v for the direction of the descent (v must satisfy the homogeneous boundary conditions of the problem in question). The next approximation is then taken to be $u^{\nu+1} = u^\nu + s^* v$, where the 'step of the descent' s^* is defined as the solution of the problem: to minimize the quadratic function of a single variable s:

$$(3.16)\qquad \min_s \Phi(s) \equiv \min_s F[u + sv].$$

Various versions of the method of minimal residuals differ in ways of constructing the function v, – a very important point which determines the efficiency of the process. It is natural to use the 'gradient descent',

that is, to take

$$v_{n,m} = -\left(\frac{\partial F}{\partial u}\right)_{n,m} = (Du - f)_{n,m}.$$

(This formula defines v at the interior points of the mesh; the values of v on the boundary are given by the homogeneous boundary conditions.) The resulting formula

$$(3.17)\qquad u^{\nu+1} = u^\nu + s^*(Du^\nu - f),$$

differs from the simple iterations (3.1) only in that the step s^* not taken beforehand on the basis of the estimates for the spectral bounds, but is determined by u^ν as a solution of (3.16). The investigation of convergence for the non-linear process (3.17)–(3.16) is considerably more complicated than for (3.1). However, these investigations have been carried out, and the convergence of the methods (3.1) and (3.17) is asymptotically the same. At the first iterations (3.17), $\|r^\nu\|$ usually decreases faster than for (3.1), but this advantage disappears relatively soon and the rate of decrease (that is, the quantity $\|r^{\nu+1}\|/\|r^\nu\|$) of the two processes is thereafter the same. One can give an example of the initial approximation u^0 for which $\|r^\nu\|$ in the process (3.17) decreases from the beginning according to the asymptotic formula (3.8).) Seidel's method is obtained from the method of minimal residuals by taking for $v_{n,m}$ the function

$$v_{n,m}^{(n^*,m^*)} = \begin{cases} 1 \text{ for } n = n^* \text{ and } m = m^*, \\ 0 \text{ at the remaining points of the mesh.} \end{cases}$$

Considering for simplicity the problem (2.1)–(2.2) and writing down the finite-difference analogue of the functional $f[u + sv]$ (3.14)

$$(3.18)\qquad h^2 \sum_{m=1}^{N-1}\sum_{n=1}^{N-1} [-\Delta(u+sv) + 2f]_{n,m}(u+sv)_{n,m},$$

we easily find that the minimum of (3.18) is attained at the point

$$(3.19)\qquad \begin{cases} s^* = \dfrac{\sum_{n=1}^{N-1}\sum_{m=1}^{N-1} (\Delta u - f)_{n,m} v_{n,m}^{(n^*,m^*)}}{\sum_{n=1}^{N-1}\sum_{m=1}^{N-1} (-\Delta v)_{n,m} v_{n,m}} = \dfrac{(\Delta u - f)_{n^*,m^*}}{4/h^2}, \\ u_{n^*,m^*}^{\nu+1} = u_{n^*,m^*}^\nu + \dfrac{h^2}{4}(\Delta u^\nu - f)_{n^*,m^*}. \end{cases}$$

A single iteration in Seidel's method consists in computing from (3.19) the values u_{n^*,m^*} successively for $n^* = 1, 2, \ldots, N-1$ and $m^* = 1, 2, \ldots, N-1$. Taking into account that the values of $u_{n,m}$ for $n < n^*$ and $m = 1, 2, \ldots, N-1$, and for $n = n^*$ and $m = 1, 2, \ldots, m^* - 1$ are already known at the time of computing u_{n^*,m^*} and should bear the superscript $\nu + 1$, we obtain the final equation

$$(3.20)\qquad u_{n,m}^{\nu+1} = \frac{1}{4}(u_{n-1,m}^{\nu+1} + u_{n,m-1}^{\nu+1} + u_{n+1,m}^\nu + u_{n,m+1}^\nu - h^2 f_{n,m})$$

$$(n, m = 1, 2, \ldots, N-1).$$

This equation can be written in a more concise form. We can represent the matrix of the system of linear algebraic equations

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{n,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)_{n,m} = f_{n,m} \tag{3.21}$$

in the form $T_H + T_B$, where T_H is the triangular matrix containing only the entries on or below the diagonal and T_B contains all the entries above the diagonal. Then Seidel's process for solving the equation $(T_H + T_B)u = f$ can be written as

$$T_H u^{\nu+1} + T_B u^{\nu} = f \quad \text{or} \quad u^{\nu+1} = T_H^{-1}(f - T_B u^{\nu}). \tag{3.22}$$

In this form the process can be generalized (at least formally) to more complicated equations without any reference to functions of the type (3.14).

In solving a boundary-value problem with the condition $\frac{\partial u}{\partial n} = 0$, say, we can act in two ways: either we start with (3.20) and then find the boundary values $u_{0,m}^{\nu+1} = u_{1,m}^{\nu+1}$ or, for example for $n = 1$, we use the formula

$$u_{1,m}^{\nu+1} = \frac{1}{3}(u_{1,m-1}^{\nu+1} + u_{1,m+1}^{\nu} + u_{2,m}^{\nu} - h^2 f_{1,m}).$$

To form some idea about the convergence of this method we make use of a device which, although not rigorous, is much exploited in numerical mathematics. Consider the process (3.20) for an unbounded problem. The eigenfunctions of the transition operator from u^{ν} to $u^{\nu+1}$ are then known, namely, they are the functions $e^{i(\xi n+\eta m)}$, where ξ and η are real parameters whose values must be restricted in some natural way. The upper estimate is obvious; $|\xi| \leqslant \pi$ and $|\eta| \leqslant \pi$ – this simply takes into account the periodicity. The lower estimate is not strict but more significant; it should to some extent compensate for the disregard of the boundary conditions.

Let R_x and R_y be the typical dimensions, the x- and y-directions, respectively, of the domain in which the problem is being solved. In the analysis of convergence we confine ourselves to the values

$$|\xi| \geqslant \frac{h\pi}{R_x}, \quad |\eta| \geqslant \frac{h\pi}{R_y} \sim \frac{\pi}{N}.$$

This is the condition for the wave-length of the smoothest of the functions in question to be of the same order as the linear size of the domain. Substituting $e^{i(\xi n+\eta m)}$ in (3.20), we compute the eigenvalues $\lambda_{\xi,\eta}$ of the transition operator corresponding to this function:

$$\lambda_{\xi,\eta} = \frac{e^{i\xi} + e^{i\eta}}{4 - e^{-i\xi} - e^{-i\eta}}. \tag{3.23}$$

The quantity $\max_{\xi,\eta} |\lambda\xi, \eta|$ enables us to make a preliminary guess about the convergence of the process. Without going into detailed analysis of (3.23),

let us compute its values, so to speak, on the boundary of the spectrum:

$$|\lambda_{\pi,\pi}| = \frac{1}{3}, \quad |\lambda_{\frac{\pi}{N},\frac{\pi}{N}}| \simeq 1 - \frac{\pi^2}{N^2}.$$

Thus, the iterations converge well for high frequencies (this estimate is fairly accurate because the high-frequency eigenfunctions are, roughly speaking, independent of the boundary conditions). The convergence is very slow for low frequencies; although the latter estimate is not very accurate, it gives an essentially correct idea.

As we have mentioned above, we can choose the direction of the descent to be $v_{n,m} = (Du^{\nu} - f)_{n,m}$ (supplementing its definition at the boundary by means of the homogeneous boundary condition). To compute the step s^* of the descent, we do not need any estimates for the bounds of the spectrum of D, but we have to compute the functions v and Dv and the two inner products $(Du - f, v)$ and (Dv, v). It is useful to note in this context that in comparing the efficiency of various iterative methods we should not look only at such parameters as 'quality' $\varkappa$ (3.8). The programmer is not interested in the number of iterations, but in the total machine time t necessary to achieve the given accuracy. Therefore apart from the 'quality' $\varkappa$, which measures the decrease of residuals with the number ν of iterations, we must also consider the machine time T in which the computer performs one iteration. Since $\nu = t/T$, the decrease of the residuals with t is given by

$$\| r^t \| \simeq \| r^0 \| e^{-\varkappa \cdot t/T}.$$

Although $\varkappa/T$ is a more objective characteristic of the process, we consider in what follows only the quantity $\varkappa$. The reason is that T depends not only on the structure of the method, but also on a number of extraneous factors, for example, the capacity of the operative memory. If the latter is $\sim 2N^2$, so that it is possible to compute and store the function $v_{n,m}$, then the iteration (3.17), including the computation of s^*, requires about twice as much time as (3.1) with the step τ predetermined: this time is mainly spent on computing $v_{n,m} = (Du - f)_{n,m}$ and $(Dv)_{n,m}$ (see (3.24)), while for (3.1) T is spent on the computation of $(Du - f)_{n,m}$. But if there are $\sim N^2$ memory cells sufficient only for the basic array $u_{n,m}$, then T is about three times greater for (3.17) than for (3.1): after computing s^* we have to compute $(Du - f)_{n,m}$ again to find $u_{n,m}^{\nu+1}$ from (3.17).

§ 4. Richardson's method

In 1910 Richardson proposed a simple and efficient method of accelerating the convergence of the simple iterations (3.1) substantially. The actual scheme for computations is no more complicated than before: the computation is to proceed according to the formula

$$u^{\nu+1} = u^{\nu} + \tau_{\nu+1}(\Delta u^{\nu} - f), \tag{4.1}$$

each iteration having its own τ_v. The gist is, of course, in the special choice of the sequence τ_v ($v = 1, 2, \ldots$). The formula for the error v^ν in the process (4.1) is an obvious generalization of (3.3):

$$v^\nu = \sum c_p \prod_{i=1}^{\nu} (1-\tau_i\lambda_p)\,\varphi^{(p)}. \tag{4.2}$$

Given a certain number ν of iterations (it is explained below how to determine ν from the required accuracy ε) we naturally demand that $\{\tau_i\}_{i=1}^{\nu}$ is a solution of the problem

$$\min_{\tau}\max_{l\leqslant\lambda\leqslant L}\Big|\prod_{i=1}^{\nu}(1-\tau_i\lambda)\Big|, \tag{4.3}$$

which immediately leads to the Chebyshev polynomial nearest to zero on $[l, L]$ and normed by the condition $T\nu(0) = 1$. Thus, one iteration by Richardson's method consists of ν iterations (4.1) with the parameters

$$\tau_i = \frac{2}{(L+l)+(L-l)\cos\frac{\pi(2i-1)}{2\nu}} \qquad (i=1, 2, \ldots, \nu). \tag{4.4}$$

The first terms $\tau_1, \tau_2, \ldots$ in the sequence (4.4) are quantities of the same order of magnitude as $1/L$; the first iterations effectively suppress those components of the error that correspond to the points of the spectrum, near its right-hand end L; the components corresponding to $\lambda \sim l$ are also suppressed, but very slowly like $e^{-\nu l/L}$. The last terms $\ldots, \tau_{\nu-1}, \tau_\nu$ of the sequence (4.4) are quantities of the order $1/l$; the last iterations intensively suppress components with $\lambda_p \sim l$ (that is, $\left|1-\frac{\lambda_p}{l}\right| \simeq 0$). But at the same time there is a strong divergence on the right-hand end of the spectrum, since there $\left|1-\frac{\lambda_p}{l}\right| \simeq \frac{L}{l}$. Altogether, in accordance with the graph of the Chebyshev polynomial the components of the error are more or less uniformly suppressed over the whole spectrum $[l, L]$. The efficiency of the method can be estimated by using the formulae for the Chebyshev polynomials. We introduce the variable $x(\lambda) = \frac{2}{L-l}\left(\lambda - \frac{l+L}{2}\right)$, and put $x_0 = x(0) \simeq -(1+2\eta)$, where $\eta = l/L$. Then

$$\prod_{i=1}^{\nu}(1-\tau_i\lambda) = \frac{(x+\sqrt{x^2-1})^\nu+(x-\sqrt{x^2-1})^\nu}{(x_0+\sqrt{x_0^2-1})^\nu+(x_0+\sqrt{x_0^2-1})^\nu}. \tag{4.5}$$

It is well-known that

$$\mu = \max_{l\leqslant\lambda\leqslant L}\Big|\prod_{i=1}^{\nu}(1-\tau_i\lambda)\Big| = \Big|\prod_{i=1}^{\nu}(1-\tau_i l)\Big| = \Big|\prod_{i=1}^{\nu}(1-\tau_i L)\Big|, \tag{4.6}$$

and thus (as $x(L) = 1$),

$$\mu = \frac{2}{(x_0+\sqrt{x_0^2-1})^\nu+(x_0-\sqrt{x_0^2-1})^\nu}.$$

As we are interested in the case $\eta \ll 1$, we have $\sqrt{x_0^2-1} \simeq \sqrt{1+4\eta+4\eta^2-1} \simeq 2\sqrt{\eta}$, and we are required to estimate the quantity

$$\mu \simeq \frac{2}{(1+2\sqrt{\eta})^\nu+(1-2\sqrt{\eta})^\nu}. \tag{4.7}$$

We do not intend to go into a full analysis of (4.7); the necessary conclusions can be obtained by considering the following three cases:

$$\begin{cases} 1.\ \nu \ll \frac{1}{2\sqrt{\eta}}: & \mu(\nu) \simeq e^{-2\frac{l}{L}\nu(\nu-1)}; \\ 2.\ \nu \simeq \frac{1}{2\sqrt{\eta}}: & \mu(\nu) \simeq \frac{2}{e^{2\nu\sqrt{\eta}}+e^{-2\nu\sqrt{\eta}}}; \\ 3.\ \nu \gg \frac{1}{2\sqrt{\eta}}: & \mu(\nu) \simeq \frac{2}{e^{2\nu\sqrt{\eta}}} \simeq e^{-2\nu\left[\sqrt{\eta}-\frac{\ln 2}{\nu}\right]}. \end{cases} \tag{4.8}$$

This shows that Richardson's method is preferable if $\nu > \frac{1}{2\sqrt{\eta}}$, when its efficiency is $\sqrt{(L/l)}$ times greater than that of the simple iterations (3.1). The solution of the example mentioned above (decreasing the initial residuals 10^5 times) on the mesh with $N = 100$ for the equation (2.1) would now require about three minutes. However, the attempts to use Richardson's method in the form in which we have described it above ended in failure; there was a complete discrepancy between computed results and theory. The fact of the matter is that the theory disregards the rounding error and a (theoretically) convergent process turned out to be even divergent. This is not difficult to explain. We have mentioned above that the evolution of the various Fourier components of the residual is of uneven character. The behaviour of the process is illustrated in Fig. 2, which shows qualitatively the changes of the component $c_\alpha^{(i)}$ corresponding to the left-hand end of the spectrum ($\lambda \sim l$), and $c_\omega^{(i)}$ corresponding to the right-hand end.

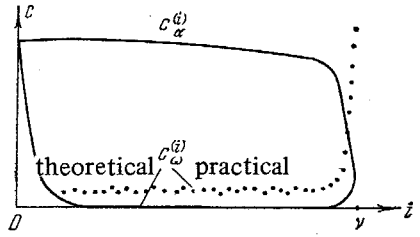


Fig. 2.

At the first iterations ($\tau \simeq 1/L$, $i = 1, 2, \ldots$), for $\lambda \sim L$ we have $(1 - \tau\lambda) \simeq 0$, and $c_\omega^{(i)}$ decreases sharply, while $c_\alpha^{(i)}$ stays nearly constant $\left(c_\alpha^{(i)} \simeq c_\alpha^{(0)}\left(1 - \frac{l}{L}\right)^i\right)$. By the end of the iteration process ($i = \ldots, \nu - 1, \nu$), $c_\alpha^{(i)}$ decreases sharply ($\tau \simeq 1/l$, $(1 - \tau\lambda) \simeq 0$ for $\lambda \sim l$); at the same time there is a strong increase of $c_\omega^{(i)}$. This coefficient is in a single iteration increased about L/l times, but this growth starts with very small values, so that the overall decrease of $c_\alpha^{(i)}$ and $c_\omega^{(i)}$ in ν iterations is the same. However, this argument ignores the rounding error $\sim |u|\varepsilon$ (where $|u|$ is the typical size of u), which implies that $c_\omega^{(i)}$ cannot stay very small. Naturally, this error is spread over the whole spectrum. Hence the sharp increase of $c_\omega^{(i)}$ at the concluding stage of the iteration cycle does not begin with its very small theoretical value, but with a much larger value determined by the rounding error. For large $\nu \simeq N \simeq 100$ this means multiplying c_ω by $\frac{L}{l} \simeq \frac{4N^2}{\pi^2}$ (that is, for $N = 100$, by 4000) in one iteration. Four such iterations are sufficient to move the rounding error from the last places of the mantissa to the first so that the theoretical analysis of convergence which we have carried out above becomes irrelevant. There are two ways of combatting this nuisance: either by using Chebyshev polynomials of low order $\nu \ll \frac{1}{\sqrt{\eta}}$ (and this leads to a marked decrease of the efficiency, see (4.8)), or by using the parameters τ_i not in their natural order (4.4), but in some different arrangement, which makes the evolution of the components of the residual over the whole of the spectrum as uniform as possible. Although empirical groping towards the second method started many years ago, the exact formulation and solution of this problem was obtained only quite recently in the paper by Lebedev and Finogenov [1]. They consider the process (4.1) and a certain permutation $\{\tau_i\}_{i=1}^{\nu}$ of the terms of the sequence (4.4), and introduce the polynomials

$$(4.9) \qquad P_k^\nu(\lambda) = \prod_{i=1}^{k} (1 - \tau_i \cdot \lambda), \quad Q_k^\nu(\lambda) = \prod_{i=k+1}^{\nu} (1 - \tau_i \cdot \lambda).$$

The first polynomial is the factor that after k iterations of the component of the residual corresponds to the eigenvalue λ; the second polynomial is the factor by which this component is to be multiplied in the remaining $\nu - k$ iterations. The problem formulated in [1] is to find permutations of the sequence (4.4) for which the quantities

$$\max_{l \leqslant \lambda \leqslant L} |P_k^\nu(\lambda)| \quad \text{and} \quad \max_{l \leqslant \lambda \leqslant L} |Q_k^\nu(\lambda)|$$

are bounded[1] for all $k = 1, 2, \ldots, \nu$ and for all ν. These permutations are found in [1] for $\nu = 2^p$; the construction turns out to be quite simple and we give it here, as it is an essential part of Richardson's method. For $p = 1$, the τ_i in (4.4) are taken in the order $\{1; 2\}$. Given the permutation $\{i_1, i_2, \ldots i_{2p}\}$, the permutation for $\nu = 2^{p+1}$ is obtained by replacing each i_j by the pair $\{i_j, 2^{p+1} + 1 - i_j\}$. So we obtain the following permutations of the τ_i in (4.4):

for $\nu = 4$: {1, 4, 2, 3},
for $\nu = 8$: {1, 8, 4, 5, 2, 7, 3, 6},
for $\nu = 16$: {1, 16, 8, 9, 4, 13, 5, 12, 2, 15, 7, 10, 3, 14, 6, 11}

etc.[2]

Let us summarize all these facts.

THEOREM 2. To decrease the norm of the residual (or the error) of the initial approximation ε^{-1} times it is sufficient to perform ν iterations according to the formula

$$u^{i+1} = u^i + \tau_{i+1}(\Delta u^i - f) \qquad (i = 1, 2, \ldots, \nu = 2^p).$$

Here $\nu \simeq \frac{1}{2}\sqrt{\frac{L}{l}} \ln \frac{1}{\varepsilon}$, and the τ_i are the terms of the sequence (4.4), ordered according to the prescription in [1].

It is worth noting that in this case the norm of the residual (error) in the iteration process decreases relatively uniformly, in accordance with the formula

$$\|r^i\| \underset{i=1, 2, \ldots \nu}{\simeq} \|r^0\| e^{-2i\sqrt{\eta}}, \quad \eta = \frac{l}{L}.$$

The scope of Richardson's method is almost as wide as that of the method of simple iterations; almost, because the convergence of simple iterations requires only the inequality $\|E + \tau D\| = \mu < 1$, which does not exclude the presence of complex eigenvalues for D. For Richardson's method this would necessitate a modification of the theory behind the choice of the parameters τ_i. However, this remark concerns general matrices D; for difference elliptic equations the appearance of complex eigenvalues is unusual. More pressing problems (from the point of view of the programmer) are connected with operators D such as, for example,

$$(4.10) \qquad D \equiv -\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + aE\right), \qquad 0 \leqslant x, y \leqslant \pi, \quad a > 2.$$

---

[1] The meaning of these conditions becomes clear if we realize that the iterations are in fact given by the formula

$$u_{n,m}^{\nu+1} = u_{n,m}^{\nu} + \tau(\Delta u^\nu - f)_{n,m} + \varepsilon_{n,m}^{\nu}.$$

The rounding error is $|\varepsilon_{n,m}^{\nu}| \sim 10^{-10}|u_{n,m}^{\nu}|$.

[2] A similar result has been obtained by Samarskii ([11], 475).

Here the spectrum is basically positive, but a few of the first eigenvalues are negative. The formula (4.3) leads naturally to the problem of constructing the polynomials that are closest to zero on the union of the two intervals $[l, l^*] \cup [l^{**}, L]$ (where $l < l^* < 0$ $l^{**} \ll L$ and l, l^* and l^{**} are numbers of the same order, while L is significantly greater). The author knows only of the simplest results in this direction [2]. We note that there is a formal device transforming any system $Du = f$ ($\det D \neq 0$) into a form suitable for the application of Richardson's method. It consists in considering the equivalent system $D^*Du = D^*f$. The operator D^*D is self-adjoint and positive; however, if we denote by l and L denote the eigenvalue of D of minimal and maximal modulus, and by l^* and L^* the bounds of the spectrum of D^*D, we obtain: $l^*/L^* = l/L|^2$, and Richardson's method gives for D^*D the decrease of the residual at the rate $e^{-2\nu|l/L|}$. Hence the universality of this device is balanced out by its too slow rate of convergence.

§ 5. Young's method

In 1954 Young proposed an iterative method, which he called the successive over-relaxation method. Its rate of convergence for such problems as (2.1)–(2.2) is the same as for Richardson's method. While Young's method is inferior to Richardson's in its scope (being only applicable to a comparatively narrow class of elliptic difference equations), it has the important advantage of being numerically stable. The evolution of the residual norm in practical work on electronic computers fits well with theoretical predictions that disregard the rounding error (of course, as long as $||r^\nu|| \geqslant |u|\varepsilon/h^2$). However, this advantage disappeared after the publication of [1], and the value of Young's method as a means of actual solution of elliptic difference equations was much diminished. We shall therefore describe it here only in general terms. A condition for the applicability of Young's method in the solution of the system of finite-difference equations

$$(Du)_{n,m} = f_{n,m} \tag{5.1}$$

is the possibility of rearranging the variables $u_{n,m}$ in such a way that the system can be written in the form

$$\begin{pmatrix} \mathscr{E}_1 & A_{12} \\ A_{21} & \mathscr{E}_2 \end{pmatrix} \begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} = \begin{Bmatrix} f_1 \\ f_2 \end{Bmatrix}. \tag{5.2}$$

(Here $\mathscr{E}_1$ and $\mathscr{E}_2$ are easily invertible, for example, diagonal matrices).

For the system

$$\begin{cases} \left(\frac{\partial^2 u}{\partial x^2}\right)_{n,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)_{n,m} = f_{n,m}, \\ u_{n,m}|_\Gamma = \varphi, \end{cases} \tag{5.3}$$

this rearranging consists in putting all the even $u_{n,m}$ (that is, those with $n + m$ even) first and all the odd $u_{n,m}$ last. Note that the presence of a mixed derivative or the use of the nine-point scheme of the fourth order for $\Delta u = f$ makes the representation (5.2) and the application of Young's scheme difficult. Formally, the iterative process [3] is written in this form:

$$\begin{cases} 1)\ u_1^{\nu+1} = (1-\omega)\, u_1^\nu + \omega \mathscr{E}_1^{-1} (f_1 - A_{12} u_2^\nu), \\ 2)\ u_2^{\nu+1} = (1-\omega)\, u_2^\nu + \omega \mathscr{E}_2^{-1} (f_2 - A_{21} u_1^{\nu+1}). \end{cases} \tag{5.4}$$

For the system (5.3) the scheme (5.4) is realized as follows:

1. First one computes for all the 'even' vertices (except, of course, on the boundary)

$$u_{n,m}^{\nu+1} = (1-\omega)\, u_{n,m}^\nu + \frac{\omega}{4} (u_{n-1,m}^\nu + u_{n,m-1}^\nu + u_{n+1,m}^\nu + u_{n,m+1}^\nu - h^2 f_{n,m}).$$

2. Then one finds for the odd vertices

$$u_{n,m}^{\nu+1} = (1-\omega)\, u_{n,m}^\nu + \frac{\omega}{4} (u_{n-1,m}^{\nu+1} + u_{n,m-1}^{\nu+1} + u_{n+1,m}^{\nu+1} + u_{n,m+1}^{\nu+1} - h^2 f_{n,m}).$$

Here $1 \leqslant \omega < 2$ is the relaxation factor, which has an essential influence on the rate of convergence. The theory behind the choice of ω is based on the computation of the spectrum of the transition operator from u^ν to $u^{\nu+1}$. The points of the spectrum of the operator (5.4) are determined from the characteristic equation

$$\det \begin{pmatrix} [\lambda - (1-\omega)]\, E & \omega \mathscr{E}_1^{-1} A_{12} \\ \lambda\omega \mathscr{E}_2^{-1} A_{21} & [\lambda - (1-\omega)]\, E \end{pmatrix} = 0. \tag{5.5}$$

Multiplying (5.5) by the matrix $\begin{pmatrix} \mathscr{E}_1 & 0 \\ 0 & \mathscr{E}_2 \end{pmatrix}$ and performing some simple transformations we obtain an equivalent equation

$$\det \begin{pmatrix} \frac{1}{\omega}\left(\sqrt{\lambda} - \frac{1-\omega}{\sqrt{\lambda}}\right) \mathscr{E}_1 & A_{12} \\ A_{21} & \frac{1}{\omega}\left(\sqrt{\lambda} - \frac{1-\omega}{\sqrt{\lambda}}\right) \mathscr{E}_2 \end{pmatrix} = 0. \tag{5.6}$$

Denoting by $\sqrt{\lambda} = z$, and w the roots of the equation

$$\det \begin{pmatrix} w\mathscr{E}_1 & A_{12} \\ A_{21} & w\mathscr{E}_2 \end{pmatrix} = 0,$$

We find the relation $\frac{1}{\omega}\left(z + \frac{1-\omega}{z}\right) = w$. Here the spectrum λ depends on the parameter ω, and its choice naturally leads to the problem

$$\min_{\omega} \max_{\lambda} |\lambda(\omega)| \tag{5.7}$$

(where the maximum is taken over the points of the spectrum). The problem (5.7) was solved in [3] for the equation (5.3); in this case $\mathscr{E}_1 = \mathscr{E}_2 = E$ and the operators A_{12} and A_{21} are determined by the formulae

$$(Au)_{n,m} = \frac{1}{4}(u_{n-1,m} + u_{n,m-1} + u_{n+1,m} + u_{n,m+1})$$

(at the interior vertices of the mesh). The eigenvalues[1] of A are $w_{p,q} = \cos\frac{p\pi}{N} + \cos\frac{q\pi}{N}$ $(p, q = 1, 2, \ldots, N-1)$ and the spectrum w lies in the interval $\left(-1+\frac{\pi^2}{N^2}, 1-\frac{\pi^2}{N^2}\right)$. The optimal value of ω is determined by

$$\omega \text{ opt} = \frac{2}{1+\sqrt{1-|w_{\max}|^2}} \simeq \frac{2}{1+\sqrt{2}\frac{\pi}{N}}. \tag{5.8}$$

The spectrum λ of the iteration operator then lies on the circle[2]

$$|\lambda| \simeq 1 - \frac{2\pi}{N} \simeq e^{-\frac{2\pi}{N}}.$$

The iterations converge uniformly, the process is numerically stable, and we can state

THEOREM 3. *For the problem* (2.1)–(2.2) *the method of successive over-relaxation* (5.4) *converges for any* ω, $1 \leqslant \omega < 2$, *and any initial approximation. For iterations with the optimal choice of* ω (5.8) *the residual norm decreases according to the formula*

$$\|r^\nu\| \simeq \|r^0\| e^{-\nu\frac{2\pi}{N}}.$$

Further developments and generalizations of the method of successive over-relaxation are treated, for example, in [4]. We omit a detailed exposition of these results for the reasons mentioned above.

§ 6. The method of alternating directions

Considerable progress in the efficiency (rate of convergence) of iterative methods for the solution of elliptic difference equations was achieved in 1955–56, when the American mathematicians Douglas, Peaceman and Rachford in [5] and [6] proposed new processes, based mainly on the "sweep" method. The sweep is an efficient way of solving special systems of linear algebraic equations that arise in the finite-difference solution of

[1] This is established by computing elementarily the result of the action of the operator on the functions $\sin\frac{p\pi n}{N} \cdot \sin\frac{q\pi m}{N}$.

[2] For $\omega \neq \omega_{\text{opt}}$ the spectrum of λ lies on an ellipse whose major semi-axis determines the rate of convergence of the iterations.

the heat equation:

$$\frac{u_n - u_n^*}{\tau} = \left(\frac{\partial^2 u}{\partial x^2}\right)_n - f_n \quad (n = 1, 2, \ldots, N-1) \tag{6.1}$$

with the boundary condition, say,

$$\alpha\left(\frac{\partial u}{\partial x}\right)_{1/2} + \beta u_0 = \varphi, \quad \alpha'\left(\frac{\partial u}{\partial x}\right)_{N-\frac{1}{2}} + \beta' u_N = \varphi'. \tag{6.2}$$

(Here u_n is the unknown and u_n^* is a known mesh function.) The obvious and formally possible procedure of solving the system (6.1)–(6.2) is to find the requisite number of solutions of the Cauchy problem for (6.1) and then to look for the solution of the boundary-value problem (6.1)–(6.2) as a linear combination of them. However, the presence of the large parameter $1/\tau$ in (6.1) makes this procedure numerically unstable.

The sweep method for (6.1)–(6.2) was proposed independently and almost at the same time in different countries; in the Soviet Union it was done in 1952–1953 by I. M. Gel'fand and O. V. Lokutsievskii.[1] It is by now widely known, and the reader should have no difficulty in finding the necessary formulae in textbooks on numerical methods. The American mathematicians proposed two constructions for iterative procedures for the equation

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{n,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)_{n,m} = f_{n,m} \tag{6.3}$$

(in a rectangular domain).

In the first the transition from u^ν to $u^{\nu+1}$ consists of two 'half-steps'.

1. First we find an intermediate function u^* from the equations

$$\frac{u^*_{n,m} - u^\nu_{n,m}}{\tau_{\nu+1}} = \left(\frac{\partial^2 u}{\partial x^2}\right)^*_{n,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)^\nu_{n,m} - f_{n,m} \quad (n, m = 1, 2, \ldots, N-1). \tag{6.4}$$

The system (6.4) splits into $N-1$ independent systems (corresponding to different values of m) of equations of the form (6.1), each of which is solved by the sweep method in the direction of x (n). We have purposely avoided specifying the boundary conditions in (6.3) and (6.4); formally they can have the general form (6.2). Of course, α and β in (6.2) cannot be completely arbitrary. This question has been studied in [7] and the restrictions on α and β stated there have a natural physical interpretation[2] and are satisfied in a very wide class of concrete problems.

2. The second half-step gives $u^{\nu+1}$ as a solution of the system

$$\frac{u^{\nu+1}_{n,m} - u^*_{n,m}}{\tau_{\nu+1}} = \left(\frac{\partial^2 u}{\partial x^2}\right)^*_{n,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)^{\nu+1}_{n,m} - f_{n,m} \quad (n, m = 1, 2, \ldots, N-1), \tag{6.5}$$

[1] This work was published only in 1962, in the Appendix to the book by Godunov and Ryaben'kii [7].

[2] Roughly speaking, we require that the heat equation $u_t = \Delta u - f$ with the boundary conditions (6.2) is stable, and that its solution tends to a finite limit as $t \to \infty$, with the limit function satisfying the equation $\Delta u = f$. In other words, the operator Δ with the (homogeneous) boundary conditions (6.2) must be negative definite.

which for differing values of n also splits into independent 'one-dimensional' equations; they are solved by the sweep method in the direction of y (m). In what follows we use a more compact form of writing this and similar algorithms, for example,

$$(6.6)\qquad 1)\ \frac{u^*-u^\nu}{\tau}=\left(\frac{\partial^2 u}{\partial x^2}\right)^*+\left(\frac{\partial^2 u}{\partial y^2}\right)^\nu-f,$$

$$2)\ \frac{u^{\nu+1}-u^*}{\tau}=\left(\frac{\partial^2 u}{\partial x^2}\right)^*+\left(\frac{\partial^2 u}{\partial y^2}\right)^{\nu+1}-f,$$

or

$$(6.7)\qquad \begin{cases} 1)\ u^*=\left(E-\tau\dfrac{\partial^2}{\partial x^2}\right)^{-1}\left[\left(E+\tau\dfrac{\partial^2}{\partial y^2}\right)u^\nu-\tau f\right], \\ 2)\ u^{\nu+1}=\left(E-\tau\dfrac{\partial^2}{\partial y^2}\right)^{-1}\left[\left(E+\tau\dfrac{\partial^2}{\partial x^2}\right)u^*-\tau f\right]. \end{cases}$$

Finally, for the evolution of the error $v^\nu = u^\nu - U$ (where U is the exact solution of (6.3)) in the iteration process it is convenient to use a form of the process with the intermediate step excluded:

$$(6.8)\qquad v^{\nu+1}=\left(E-\tau\frac{\partial^2}{\partial y^2}\right)^{-1}\left(E+\tau\frac{\partial^2}{\partial x^2}\right)\left(E-\tau\frac{\partial^2}{\partial x^2}\right)^{-1}\left(E+\tau\frac{\partial^2}{\partial y^2}\right)v^\nu.$$

It is worth emphasizing that the (finite-dimensional) operator $\left(E-\tau\dfrac{\partial^2}{\partial x^2}\right)^{-1}$ in (6.7) and (6.8) is realized in the form of an algorithm for solving problems of the type (6.1)–(6.2), rather than as a matrix. The second algorithm of the American authors, although it leads to a process with a slightly slower rate of convergence, nevertheless has an important advantage: it can be generalized to equations in three (and more) independent variables. Its scheme (which we write down at once for a three-dimensional equation) is the following:

$$(6.9)\qquad \begin{cases} 1)\ \dfrac{u^*-u^\nu}{\tau_{\nu+1}}=\left(\dfrac{\partial^2 u}{\partial x^2}\right)^*+\left(\dfrac{\partial^2 u}{\partial y^2}\right)^\nu+\left(\dfrac{\partial^2 u}{\partial z^2}\right)^\nu-f, \\ 2)\ \dfrac{u^{**}-u^*}{\tau_{\nu+1}}=\left(\dfrac{\partial^2 u}{\partial y^2}\right)^{**}-\left(\dfrac{\partial^2 u}{\partial y^2}\right)^\nu, \\ 3)\ \dfrac{u^{\nu+1}-u^{**}}{\tau_{\nu+1}}=\left(\dfrac{\partial^2 u}{\partial z^2}\right)^{\nu+1}-\left(\dfrac{\partial^2 u}{\partial z^2}\right)^\nu. \end{cases}$$

According to the scheme (6.9) we first use the sweep method in x to determine the intermediate function u^* from the system split into independent one-dimensional equations (for distinct values of y and z). Next we compute the second intermediate function u^{**} by sweeps in y, and finally, the function $u^{\nu+1}$ by sweeps in z. In the form (6.9) the realization of the method requires $2N^3$ memory cells since both u^ν and u^* are needed to determine u^{**}, and u^ν to determine $u^{\nu+1} - u^{**}$. In some cases this makes the use of these algorithms on electronic computers difficult, and so it

makes sense to give an algorithm of a different form, by writing (6.9) as

$$(6.10)\qquad \begin{cases} 1)\ \left(E-\tau\dfrac{\partial^2}{\partial x^2}\right)u^*=\left[E+\tau\left(\dfrac{\partial^2}{\partial y^2}+\dfrac{\partial^2}{\partial z^2}\right)\right]u^\nu-\tau f, \\ 2)\ \left(E-\tau\dfrac{\partial^2}{\partial y^2}\right)u^{**}=u^*-\tau\left(\dfrac{\partial^2 u}{\partial y^2}\right)^\nu, \\ 3)\ \left(E-\tau\dfrac{\partial^2}{\partial z^2}\right)u^{\nu+1}=u^{**}-\tau\left(\dfrac{\partial^2 u}{\partial z^2}\right)^\nu. \end{cases}$$

Applying $\left(E-\tau\dfrac{\partial^2}{\partial y^2}\right)$ to the last equation in (6.10) and replacing $\left(E-\tau\dfrac{\partial^2}{\partial y^2}\right)$ u^{**} by the right-hand side of the second equation, we obtain

$$\left(E-\tau\frac{\partial^2}{\partial y^2}\right)\left(E-\frac{\partial^2}{\partial z^2}\right)u^{\nu+1}=u^*-\tau\left(\frac{\partial^2 u}{\partial y^2}\right)^\nu-\tau\left(E-\frac{\partial^2}{\partial y^2}\right)\left(\frac{\partial^2}{\partial z^2}\right)u^\nu.$$

Here, in turn, we apply $\left(E-\tau\dfrac{\partial^2}{\partial x^2}\right)$, and replace $\left(E-\tau\dfrac{\partial^2}{\partial x^2}\right)u^*$ by the right-hand side of the first equation in (6.10) we obtain[1]

$$(6.11)\qquad \left(E-\tau\frac{\partial^2}{\partial x^2}\right)\left(E-\tau\frac{\partial^2}{\partial y^2}\right)\left(E-\tau\frac{\partial^2}{\partial z^2}\right)u^{\nu+1}=Ru^\nu-\tau f,$$

where

$$(6.12)\qquad R\equiv E+\tau\left(\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+\tau\left(E-\tau\frac{\partial^2}{\partial x^2}\right)\frac{\partial^2}{\partial y^2}-$$
$$-\tau\left(E-\tau\frac{\partial^2}{\partial x^2}\right)\left(E-\tau\frac{\partial^2}{\partial y^2}\right)\frac{\partial^2}{\partial z^2}.$$

In practice, (6.11) is realized in three steps:

1) u^* is found from $\left(E-\tau\dfrac{\partial^2}{\partial x^2}\right)u^*=Ru^\nu-\tau f$;

2) then u^{**} is found from $\left(E-\tau\dfrac{\partial^2}{\partial y^2}\right)u^{**}=u^*$;

3) finally, $u^{\nu+1}$ from: $\left(E-\tau\dfrac{\partial^2}{\partial z^2}\right)u^{\nu+1}=u^{**}$.

Computation according to this scheme does not require the additional memory: as $u^*(u^{**}, u^{\nu+1})$ is computed, we may "forget" $u^\nu(u^*, u^{**})$. However, the computation of Ru for $\tau \simeq 1$ involves an error of

$O\left(\dfrac{\varepsilon\tau^3|u|}{h^6}\right)$ (and such τ, as we shall see, are necessary.)

---

[1] The difference operator R approximates a degenerate differential operator of the sixth order. The structure of R is such that if u is given at all the mesh points, then Ru is determined at all the interior points. The operator $B=\left(E-\tau\dfrac{\partial^2}{\partial x^2}\right)\left(E-\tau\dfrac{\partial^2}{\partial y^2}\right)\left(E-\tau\dfrac{\partial^2}{\partial z^2}\right)$ is also degenerate and of the sixth order. To determine $u^{\nu+1}$ uniquely from the equation $Bu^{\nu+1}=Ru^\nu-\tau f$, which is defined at all the interior mesh points, it is enough to use the boundary conditions (6.2) of the second order equation (6.3).

Choice of iteration parameters. The approximate theory. The choice of the parameters τ_ν is the essential factor determining the efficiency of the iterative processes (6.7) and (6.9). The authors of the method proposed a simple and elegant way of constructing sequences τ_ν, which we describe here for the iteration process (6.9) of solving, for example, the first boundary-value problem for Poisson's equation in the $(\pi \times \pi \times \pi)$-cube. As a matter of fact, the theory carries over almost without any changes to the more general problem:

$$(6.13) \qquad \frac{\partial}{\partial x} a(x) \frac{\partial u}{\partial x} + \frac{\partial}{\partial y} b(y) \frac{\partial u}{\partial y} + \frac{\partial}{\partial z} c(z) \frac{\partial u}{\partial z} = f(x, y, z)$$

with the boundary conditions

$$(6.14) \qquad \alpha \frac{\partial u}{\partial n} + \beta u = \varphi.$$

Of course, $a(x)$, $b(y)$ and $c(z)$ are assumed to be positive and α and β are constant on each face of the cube; furthermore, we assume that the difference operator

$$-\frac{\partial}{\partial x} a(x) \frac{\partial}{\partial x}, \quad 0 \leqslant x \leqslant \pi,$$

is strictly positive on the set of the mesh functions satisfying the boundary conditions

$$\alpha_0' \frac{\partial u}{\partial x} + \beta_0' u = 0 \Big|_{x=0}, \quad \alpha_1' \frac{\partial u}{\partial x} + \beta_1' u \Big|_{x=\pi} = 0$$

(α_0' and β_0' (α_1' and β_1') being the constants in (6.14) that correspond to the faces $x = 0$ and $x = \pi$). Let $\psi_n^{(p)}$ denote the eigenfunctions of this operator, λ_p' the corresponding eigenvalues, and l' and L' the bounds of the spectrum (so that $0 < l' \leqslant \lambda_p' \leqslant L'$). We make similar assumptions about

$-\frac{\partial}{\partial y} b(y) \frac{\partial}{\partial y}$ and $-\frac{\partial}{\partial z} c(z) \frac{\partial}{\partial z}$ and use the notation $\chi_m^{(q)}$, $l'' \leqslant \lambda_q'' \leqslant L''$ and

$\rho_k^{(r)}$, $l''' < \lambda_r''' \leqslant L'''$ for the corresponding eigenfunctions and eigenvalues. In the space of the mesh functions satisfying the homogeneous boundary conditions on the $(\pi \times \pi \times \pi)$-cube the operators $-\frac{\partial}{\partial x} a \frac{\partial}{\partial x}$, $-\frac{\partial}{\partial y} b \frac{\partial}{\partial y}$ and $-\frac{\partial}{\partial z} c \frac{\partial}{\partial z}$ have common eigenfunctions $\varphi_{n,m,k}^{(p,q,r)} = \psi_n^{(p)} \chi_m^{(q)} \rho_k^{(r)}$ with the eigenvalues λ_p', λ_q'' and λ_r''', respectively. From (6.11) we conclude that the error v^ν in the iteration process evolves by the formula

$$(6.15) \qquad \left(E - \tau_{\nu+1} \frac{\partial^2}{\partial x^2}\right)\left(E - \tau_{\nu+1} \frac{\partial^2}{\partial y^2}\right)\left(E - \tau_{\nu+1} \frac{\partial^2}{\partial z^2}\right) v^{\nu+1} = R v^\nu.$$

The function v satisfies the homogeneous boundary conditions (6.2) and

can be expanded as a Fourier series in φ:

$$v^\nu = \sum_{p,q,r} c_{p,q,r}^\nu \varphi^{(p,q,r)}, \quad \| v^\nu \| = \Big[\sum_{p,q,r} (c_{p,q,r}^\nu)^2 \Big]^{1/2}.$$

By (6.15) we find

$$v^{\nu+1} = \sum_{p,q,r} c_{p,q,r}^\nu Q(\tau_{\nu+1}; \lambda_p', \lambda_q'', \lambda_r''') \varphi^{(p,q,r)},$$

where

$$(6.16) \quad Q(\tau; \lambda', \lambda'', \lambda''') \equiv$$
$$\equiv \frac{1 - \tau(\lambda'' + \lambda''') + \tau(1 + \tau\lambda')\lambda'' + \tau(1 + \tau\lambda')(1 + \tau\lambda'')\lambda'''}{(1 + \tau\lambda')(1 + \tau\lambda'')(1 + \tau\lambda''')} =$$
$$= \frac{1 + \tau^2(\lambda'\lambda'' + \lambda''\lambda''' + \lambda'''\lambda') + \tau^3\lambda'\lambda''\lambda'''}{1 + \tau^2(\lambda'\lambda'' + \lambda''\lambda''' + \lambda'''\lambda') + \tau^3\lambda'\lambda''\lambda''' + \tau(\lambda' + \lambda'' + \lambda''')}.$$

(Note that the lack of symmetry in the process (6.9) with respect to the variables x, y and z is only apparent, since λ', λ'' and λ''' occur in Q symmetrically.)

For every $\tau < 0$ and $\lambda > 0$ we have $|Q| < 1$, and the process converges. Further, it is obvious that

$$v^\nu = \sum_{p,q,r} c_{p,q,r}^0 \prod_{i=1}^{\nu} Q(\tau_i; \lambda_p', \lambda_q'', \lambda_r''') \varphi^{(p,q,r)},$$

and, fixing for a while the number of iterations ν, it is natural to choose the sequence $\{\tau_i\}_{i=1}^\nu$ by solving the problem

$$\min_\tau \max_\lambda \prod_{i=1}^{\nu} Q(\tau_i; \lambda', \lambda'', \lambda''')$$

(max is taken over $[l', L'] \times [l'', L''] \times [l''', L''']$). Let $l = \min(l', l'', l''')$, $L = \max(L', L'', L''')$ and note that it is sufficient to estimate $\prod_i Q$ on the diagonal[1] $\lambda' = \lambda'' = \lambda''' = \lambda$, $l \leqslant \lambda \leqslant L$.
Thus, we consider

$$\prod_{i=1}^{\nu} \left[1 - \frac{3\tau_i\lambda}{(1 + \tau_i\lambda)^3}\right] = \prod_{i=1}^{\nu} Q(\tau_i \cdot \lambda),$$

where $Q(\xi) = 1 - \frac{3\xi}{(1+\xi)^3}$. The graph of $Q(\xi)$ is shown in Fig. 3. For $\frac{5}{9} < \theta < 1$, we say that the set $\{\xi: Q(\xi) \leqslant \theta\}$ is the θ-interval and we denote its bounds by $\Lambda(\theta)$ and $\Pi(\theta)$ (see Fig. 3); the bounds of the θ-interval of $Q(\tau\lambda)$ are then $\Lambda(\theta)/\tau$ and $\Pi(\theta)/\tau$.

---

[1] $Q(\tau, \lambda', \lambda'', \lambda''') = \left[1 + \frac{\tau(\lambda' + \lambda'' + \lambda''')}{P(\tau, \lambda', \lambda'', \lambda''')}\right]^{-1}$, where $P(\tau, \lambda', \lambda'', \lambda''') = 1 + \tau^2(\lambda'\lambda'' + \lambda''\lambda''' + \lambda'''\lambda') + \tau^3\lambda'\lambda''\lambda'''$. Note that Q attains its maximum on the plane $\lambda' + \lambda'' + \lambda''' = \text{const}$ at the same point as P, that is, for $\lambda' = \lambda'' = \lambda'''$. Every point of the spectrum $l \leqslant \lambda', \lambda'', \lambda''' \leqslant L$ is contained in some plane $\lambda_1 + \lambda_2 + \lambda_3 = \text{const} = \lambda' + \lambda'' + \lambda'''$, which intersect the diagonal in $(\lambda, \lambda, \lambda)$, where $3\lambda = \lambda' + \lambda'' + \lambda'''$ and $l \leqslant \lambda \leqslant L$.

We now choose τ_1 so that the left end-point of the first θ-interval coincides with the left end-point l of the spectrum, that is, $\Lambda(\theta)/\tau_1 = l$, or $\tau_1 = \Lambda(\theta)/l$. Then $Q(\tau_1\ \lambda) \leqslant \theta$ and, therefore, $\prod_{i=1}^{\nu} Q(\tau_i\lambda) \leqslant \theta$ on the whole interval $\left[l,\ \frac{\Pi(\theta)}{\tau_1} = l\frac{\Pi(\theta)}{\Lambda(\theta)}\right]$. Next, we choose

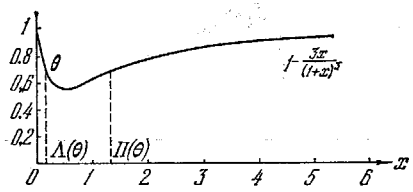


Fig. 3.

τ_2 so that the left end-point of the second interval coincides with the right end-point of the first:

$$\frac{\Lambda(\theta)}{\tau_2} = \frac{\Pi(\theta)}{\tau_1} = l\frac{\Pi(\theta)}{\Lambda(\theta)}, \text{ that is, } \tau_2 = \tau_1\frac{\Lambda(\theta)}{\Pi(\theta)}.$$

Now we can claim that $\prod_{i=1}^{\nu} Q(\tau_i\lambda) < \theta$ on $\left[l,\ \frac{\Pi(\theta)}{\tau_2} = l\left(\frac{\Pi}{\Lambda}\right)^2\right]$. Continuing in this manner, we obtain the sequence

$$\tau_1 = \frac{\Lambda(\theta)}{l},\ \tau_2 = \tau_1\frac{\Lambda(\theta)}{\Pi(\theta)},\ \ldots,\ \tau_\nu = \tau_{\nu-1}\frac{\Lambda(\theta)}{\Pi(\theta)},\ \text{and after } \nu \text{ iterations}$$

$$\prod_{i=1}^{\nu} Q(\tau_i\lambda) < \theta \quad \text{при} \quad l \leqslant \lambda \leqslant l\left[\frac{\Pi(\theta)}{\Lambda(\theta)}\right]^{\nu}.$$

The number ν is determined by the condition

$$l\left[\frac{\Pi(\theta)}{\Lambda(\theta)}\right]^{\nu} \geqslant L \geqslant l\left[\frac{\Pi(\theta)}{\Lambda(\theta)}\right]^{\nu-1}.$$

Thus, by performing ν iterations according to the scheme (6.9), we diminish the norm of the error θ^{-1} times; we continue the computation with the same sequence $\tau_1, \tau_2, \ldots, \tau_\gamma$ until we achieve the desired accuracy. On average, one iteration diminishes the error $\theta^{-1/\nu}$ times, and it is natural to choose the parameter θ so that $\theta^{1/\nu}$ is minimized. This is easily done if we use a table for $\theta^{1/\nu}$, such as the table given here (for the problem $\Delta u = f$ in a cube, with $N = 100$, boundary conditions of the first kind, $l = 1$ and $L \sim 4000$). We use the notation

θ	$\Lambda(\theta)$	$\Pi(\theta)$	$\nu(\theta)$	$\gamma(\theta)$	ν
0.6	0.25	0.90	6.5	0.079	7
0.65	0.20	1.16	4.7	0.091	5
0.70	0.16	1.45	3.8	0.095	4
0.75	0.12	1.80	3.1	0.096	3
0.80	0.10	2.2	2.7	0.083	3
0.90	0.04	4.0	1.8	0.050	2

$$\nu(\theta) = \ln\left(\frac{L}{l}\right)\Big/\ln\frac{\Pi(\theta)}{\Lambda(\theta)},\quad \gamma(\theta) = -\frac{\ln\theta}{\nu(\theta)}$$

(for the choice of ν, these functions need only be computed with a low accuracy). It is clear that we have to take a cycle with $\nu = 3$ or 4 ($\theta = 0.75$ or 0.7). The iteration parameters are as follows:

$\nu = 3$: $\tau_1 = \frac{\Lambda}{l} = 0.12$; $\tau_2 = \tau_1 \cdot 0.068 = 0.008$; $\tau_3 = 0.00053$;

$\nu = 4$: $\tau_1 = 0{,}16$; $\tau_2 = 0.0177$; $\tau_3 = 0.00195$; $\tau_4 = 0.000215$.

The error in the iteration process decreases at the rate

$$\|v^t\| \simeq \|v^0\| e^{-0.095t/T_3},$$

where T_3 is the time needed for the three steps of one iteration (6.9). In the same situation, Richardson's method gives the rate

$$\|v^t\| \simeq \|v^0\| e^{-0.031t/T_1}.$$

Taking into account that $T_3 \geqslant 3T_1$, that the solution of the three-dimensional equation with $N = 100$ requires 10^6 memory cells, which is at present unrealistic, and, finally, that the efficiency of Richardson's method grows with decreasing N faster than the efficiency of (6.9), we conclude that for the equation in (x, y, z) Richardson's method is at present more efficient than the alternating direction method. True, one can imagine a situation where this is not so: consider, for example, the equation

$$u_{xx} + u_{yy} + u_{zz} + au = f$$

where $a > 0$ is such that the first eigenvalue λ_1 satisfies $0 < \lambda_1 \ll 1$.) The optimal value of θ is determined as the minimal point of the function $\exp\left\{-\frac{\ln\theta\,[\ln\Pi(\theta) - \ln\Lambda(\theta)]}{\ln(L/l)}\right\}$. Without solving this problem we observe that the choice of θ is independent of L/l, that is of the concrete form of the system. We put $\varkappa = \max_\theta\left[\ln\theta^{-1}\cdot\ln\frac{\Pi(\theta)}{\Lambda(\theta)}\right]$. The theorem below is stated in a general form, which takes into account only those properties of the system that are actually used in the preceding analysis.

THEOREM 4. *To solve the system*

$$Du \equiv D_1u + D_2u + D_3u = f,$$

where D_i are commuting negative-definite operators whose spectra lie in the interval $[-L, -l]$, the iterative process

$$\frac{u^* - u^\nu}{\tau_{\nu+1}} = D_1u^* + (D_2 + D_3)u^\nu - f,$$

$$\frac{u^{**} - u^*}{\tau_{\nu+1}} = D_2u^{**} - D_2u^\nu,$$

$$\frac{u^{\nu+1} - u^{**}}{\tau_{\nu+1}} = D_3u^{\nu+1} - D_3u^\nu$$

converges for any $\tau > 0$. *With the special choice of a* ν*-periodic sequence of parameters* τ_i *the norm of the error (or residual) decreases on average according to the formula*

$$\| v^\nu \| \simeq \| v^0 \| \exp\left(-\frac{\varkappa \nu}{\ln L/l}\right), \quad \| r^\nu \| \simeq \| r^0 \| \exp\left(-\frac{\varkappa \cdot \nu}{\ln L/l}\right),$$
$$\varkappa \simeq 0{\cdot}8.$$

Apart from the stated properties of the operators D_i, the usefulness of this process depends mainly on the efficient procedure for solving a system of equations of the form

$$(E - \tau D_i)u = \varphi.$$

If in estimating $\prod_i Q$ we take into account only one factor for every point $(\lambda', \lambda'', \lambda''')$, we obtain an upper estimate of $\|v^\nu\|$, which is not, however, too rough. Fig. 4 shows the functions $Q(\tau_2\lambda)$ and $Q(\tau_3\lambda)$ in the example with $N = 100$ ($\nu = 3$, $\theta = 0.75$; $Q(\tau_1\lambda)$ cannot be shown on this scale). It is clear that $Q(\tau_1\lambda)Q(\tau_2\lambda)$ differs little from 1 for $500 \leqslant \lambda \leqslant 4000 = L$.

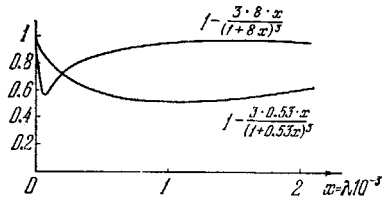


Fig. 4.

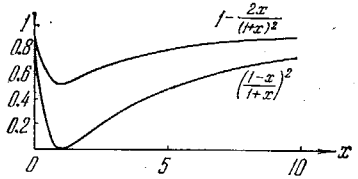


Fig. 5.

The analysis of convergence and the theory behind the choice of the sequence of iteration parameters $\{\tau_i\}$ for the first procedure are virtually the same as above. Here we determine ν and $\{\tau_1, \tau_2, \ldots, \tau_\nu\}$ from the condition

$$(6.17) \quad \min_\tau \max_{\lambda', \lambda''} \left| \prod_{i=1}^{\nu} \frac{1-\tau_i\lambda'}{1+\tau_i\lambda'} \frac{1-\tau_i\lambda''}{1+\tau_i\lambda''} \right| = \min_\tau \max_\lambda \prod_{i=1}^{\nu} \left(\frac{1-\tau_i\lambda}{1+\tau_i\lambda}\right)^2.$$

The function $Q(\xi) = \left(\frac{1-\xi}{1+\xi}\right)^2$ is shown in Fig. 5 together with the function $1 - \frac{2\xi}{(1+\xi)^2}$ which corresponds to the process

$$\frac{u^* - u^\nu}{\tau} = D_1 u^* + D_2 u^\nu - f, \quad \frac{u^{\nu+1} - u^*}{\tau} = D_2 (u^{\nu+1} - u^\nu).$$

Clearly, the first process is more efficient; the choice of the optimal ν is

illustrated (for the same example with $N = 100$) in Table 2.[1]

θ	$\Lambda(\theta)$	$\Pi(\theta)$	$\nu(\theta)$	$\gamma(\theta)$	ν
0.010	0.84	1.23	21.9	0.210	22
0.040	0.69	1.70	9.2	0.350	9
0.090	0.56	1.88	6.9	0.360	7
0.123	0.48	2.10	5.6	0.375	6
0.160	0.42	2.36	4.8	0.384	5
0.201	0.36	2.64	4.2	0.382	4
0.250	0.32	3.0	3.7	0.375	4
0.360	0.25	4.0	3.0	0.340	3

We can state

THEOREM 5. *To solve the system*

$$Du \equiv D_1 u + D_2 u = f,$$

where the D_i *are commuting negative-definite operators whose spectra lie in the interval* $[-L, -l]$ *the iteration process*

$$\frac{u^* - u^\nu}{\tau_{\nu+1}} = D_1 u^* + D_2 u^\nu - f,$$
$$\frac{u^{\nu+1} - u^*}{\tau_{\nu+1}} = D_1 u^* + D_2 u^{\nu+1} - f$$

converges for any $\tau > 0$. *The special choice of a* ν*-periodic sequence of parameters* $\{\tau_1, \tau_2, \ldots, \tau_\nu\}$ *gives an iteration process in which the norm of the error (residual) decreases in accordance with the formula*

$$\| v^\nu \| \simeq \| v^0 \| \exp\left(-\frac{\varkappa \nu}{\ln L/l}\right), \quad \varkappa = \max_\theta \ln \theta^{-1} \cdot \ln \frac{\Pi(\theta)}{\Lambda(\theta)} \simeq 3.2.$$

By using this process, the problem with $N = 100$ and the accuracy $\varepsilon = 10^{-5}$ is solved in 30 iterations.[2] Above we have simplified the theory somewhat, by neglecting the differences between the quantities l' and l'', and L' and L'' and considering the minimax problem (6.17) not on the rectangle $[l', L'] \times [l'', L'']$, but on the square $[l, L] \times [l, L]$, which contains it. If there is a large difference in the position of the spectra, this results in a loss of efficiency. A more accurate theory for the choice of τ can be obtained by applying a fractional-linear transformation. The matter will be explained in the exposition of the exact theory of the choice of parameters (for the two-dimensional problem).

[1] By taking at each point of the spectrum λ only one factor we increase the value of $\prod_i Q$. By computing $\prod_i Q$ at $\lambda = l$ it can be shown that the coefficient of the first eigenfunction in the expansion of v^ν decreases not faster than $e^{-0{\cdot}45\nu}$.

[2] The solution equations of the type of $(E - \tau\frac{\partial^2}{\partial x^2})u = \varphi$ (with the corresponding boundary conditions) by the sweep method requires $O(N^p)$ operations (where p is the number of independent variables). Therefore the number of operations that are necessary to decrease $\| v^0 \|$, $\| r^0 \|$ ε^{-1} times is estimated by the quantity $O(N^p \ln N \cdot \ln\frac{1}{\varepsilon})$.

§ 7. The choice of the optimal sequence of iteration parameters. Wachspress' theory

In [8] Wachspress proposes for solving the system of difference equations

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{n,m}+\left(\frac{\partial^2 u}{\partial y^2}\right)_{n,m}=f_{n,m}$$

in a rectangular domain (with boundary conditions of the third kind) the iteration process

$$(7.1)\qquad \begin{cases} \dfrac{u^*-u^\nu}{\tau'_{\nu+1}}=\left(\dfrac{\partial^2 u}{\partial x^2}\right)^*+\left(\dfrac{\partial^2 u}{\partial y^2}\right)^\nu-f, \\ \dfrac{u^{\nu+1}-u^*}{\tau''_{\nu+1}}=\left(\dfrac{\partial^2 u}{\partial x^2}\right)^*+\left(\dfrac{\partial^2 u}{\partial y^2}\right)^{\nu+1}-f. \end{cases}$$

The choice of the sequences τ'_ν and τ''_ν of iteration parameters is based in [8] on the exact solution of a Chebyshev-type problem for the function $Q(\lambda, \{\tau', \tau''\})$. This result is important both from the practical point of view (this choice of τ appreciably increases the efficiency of (7.1)) and also from the theoretical point of view because it explains the limitations of the method of alternating directions. The theory of the choice of τ' and τ'' is much simpler for $\nu = 2^s$ than for the general ν; as the requirement that $\nu = 2^s$ does not cause any inconvenience in practice, we consider only this case. The exposition that follows is in essence the same as in Wachspress' original work, with some, purely editorial, simplifications.

We note that the whole theory applies to the equation

$$D_1 u + D_2 u = f$$

provided that: 1) D_1 and D_2 are negative, bounded, self-adjoint and commuting operators, and 2) the solution of the system of the form

$$(E - \tau D_i)u = \varphi \quad (i = 1, 2)$$

is considerably easier than that of the original problem. Below we clarify what class of important problems leads to such operators. Just as in [8], the only concrete features of the operators we use are the spectral bounds for $-D_1$ and $-D_2$:

$$0 < l' \leqslant \lambda' \leqslant L', \quad 0 < l'' \leqslant \lambda'' \leqslant L''.$$

The problem of choosing τ' and τ'' can be stated as follows: to find a sequence $\{\tau\} = \{\tau'_1, \tau''_1; \tau'_2, \tau''_2; \ldots; \tau'_\nu, \tau''_\nu\}$ that gives

$$(7.2)\qquad \min_{\tau}\max_{\lambda', \lambda''}\left|\prod_{i=1}^{\nu}\frac{1-\tau''_i\lambda'}{1+\tau'_i\lambda'}\cdot\frac{1-\tau'_i\lambda''}{1+\tau''_i\lambda''}\right|,$$

where $\nu = 2^s$ and the maximum is taken over $[l', L'] \times [l'', L'']$.

LEMMA 1. *If we replace the variables λ', λ'', τ', τ'' in (7.2) by new variables μ', μ'', ξ', ξ'' connected with them by a fractional-linear transformation*

$$\lambda'=\frac{\mu'-p}{q-r\cdot\mu'}, \quad \lambda''=\frac{\mu''+p}{q+r\cdot\mu''}, \quad \xi'=\frac{r+\tau''}{q+p\tau''}, \quad \xi''=\frac{\tau'-r}{q-p\cdot\tau'}$$

(where p, q and r are the parameters of the transformation), we can reduce (7.2) to a problem with identical domains for the new variables $\eta \leqslant \mu', \mu'' \leqslant 1$:

$$(7.2^*)\qquad \min_{\xi}\max_{\eta\leqslant\mu', \mu''\leqslant 1}\left|\prod_{i=1}^{\nu}\frac{1-\xi''_i\mu'}{1+\xi'_i\mu'}\cdot\frac{1-\xi'_i\mu''}{1+\xi''_i\mu''}\right|.$$

The proof is by a simple verification and is therefore omitted. However, later we need the formulae for computing the parameters p, q, r and η of the transformation; these are obtained from the system of equations $\mu'(l') = \eta$, $\mu'(L') = 1$, $\mu''(l'') = \eta$, $\mu''(L'') = 1$. We set

$$(7.3)\qquad \begin{cases} A=\dfrac{L'+L''}{2L'L''}, \quad B=\dfrac{L'-L''}{2L'L''}, \quad a=\dfrac{1}{A}\left(B+\dfrac{1}{l'}\right), \\ b=\dfrac{1}{A}\left(B-\dfrac{1}{l''}\right), \quad c=\dfrac{1-ab}{a-b}. \end{cases}$$

Then

$$(7.4)\qquad \begin{cases} \eta=c-\sqrt{c^2-1}, & p=\dfrac{\eta(a-b)-2}{a+b}. \\ q=A+\eta B, & r=B+\eta A. \end{cases}$$

To solve the problem (7.2), we find it convenient to modify its statement first in an obvious way.

LEMMA 2. *The sequence $\{t\}$ that realizes*

$$(7.5)\qquad \min_{\{t\}}\max_{\eta\leqslant\lambda', \lambda''\leqslant 1}\left|\prod_{i=1}^{\nu}\frac{t'_i-\lambda'}{t'_i+\lambda'}\cdot\frac{t''_i-\lambda''}{t''_i+\lambda''}\right|$$

exists, is unique, and consists of the pairs of equal numbers $t' = t''$, $i = 1, 2, \ldots, \nu$. (Here $\tau'_i = 1/t_i$, $\tau''_i = 1/t_i$).

The proof of this lemma and the exposition of the corresponding minimax theory for (7.5) has no direct bearing on the computation of τ and is deferred until later.

LEMMA 3. *Let $t_1 < t_2 < \ldots < t_\nu$ be an optimal sequence. Then $\eta/t_1, \eta/t_2, \ldots, \eta/t_\nu$ is also an optimal sequence (the same, but in reversed order).*

The proof follows from the identity

$$\left|\frac{t-\lambda}{t+\lambda}\right|=\left|\frac{\eta/t-\eta/\lambda}{\eta/t+\eta/\lambda}\right|=\left|\frac{t'-\lambda'}{t'+\lambda'}\right|$$

(where $t' = \eta/t$, $\lambda' = \eta/\lambda$ and $\lambda' \in [\eta, 1]$ for $\lambda \in [\eta, 1]$) and from the uniqueness of $\{t\}$.

We now rewrite the product for $\nu = 2^s$:

$$\prod_{i=1}^{2^s} \frac{t_i-\lambda}{t_i+\lambda} = \prod_{i=1}^{2^{s-1}} \frac{t_{\nu+1-i}-\lambda}{t_{\nu+1-i}+\lambda}\frac{t_i-\lambda}{t_i+\lambda} = \prod_{i=1}^{2^{s-1}} \frac{t_i-\lambda}{t_i+\lambda}\cdot\frac{\eta/t_i-\lambda}{\eta/t_i+\lambda}$$

(by Lemma 3, $t_{\nu+1-i} = \eta/t_i$). Further,

$$\frac{t-\lambda}{t+\lambda}\frac{\eta/t-\lambda}{\eta/t+\lambda} \equiv \frac{\frac{1}{1+\eta}\left(\lambda+\frac{\eta}{\lambda}\right)-\frac{1}{1+\eta}\left(t+\frac{\eta}{t}\right)}{\frac{1}{1+\eta}\left(\lambda+\frac{\eta}{\lambda}\right)+\frac{1}{1+\eta}\left(t+\frac{\eta}{t}\right)}.$$

Introducing new variables $\lambda' = \frac{1}{1+\eta}\left(\lambda+\frac{\eta}{\lambda}\right)$, $t' = \frac{1}{1+\eta}\left(t+\frac{\eta}{t}\right)$ and taking into account that $\lambda' \in [\eta', 1]$ for $\lambda \in [\eta, 1]$, where $\eta' = \frac{2\sqrt{\eta}}{1+\eta}$, we obtain

$$\max_{\eta\leqslant\lambda\leqslant 1}\left|\prod_{i=1}^{2^s}\frac{t_i-\lambda}{t_i+\lambda}\right| = \max_{\eta'\leqslant\lambda\leqslant 1}\left|\prod_{i=1}^{2^{s-1}}\frac{t_i'-\lambda}{t_i'+\lambda}\right|.$$

So we have proved

LEMMA 4. *If $\{t_1, t_2, \ldots, t_{2^s}\}$ is the optimal sequence of order 2^s on the interval $[\eta, 1]$, then $t_i' = \frac{1}{1+\eta}\left(t_i+\frac{\eta}{t_i}\right)$ $(i=1, 2, \ldots, 2^{s-1})$ is the optimal sequence of order 2^{s-1} on $[\eta', 1]$.*

LEMMA 5. *If t_i', $(i=1, 2, \ldots, 2^{s-1})$, is the optimal sequence of order 2^{s-1} on the interval $[\eta', 1]$ and $\eta' = \frac{2\sqrt{\eta}}{1+\eta}$, then*

$$(7.6)\qquad \begin{cases} t_i = \frac{1+\eta}{2}t_i' - \sqrt{\left(\frac{1+\eta}{2}t_i'\right)^2-\eta}, \\ t_{2^s+1-i} = \frac{1+\eta}{2}t_i' + \sqrt{\left(\frac{1+\eta}{2}t_i'\right)^2-\eta} \quad (i=1, 2, \ldots, 2^{s-1}) \end{cases}$$

is the optimal sequence of the order 2^s on $[\eta, 1]$.

Next, we reduce the problem of finding the optimal sequence of order 2^{s-1} to the problem of finding the optimal sequence of order 2^{s-2} etc., down to order 2^0. The latter problem is readily solved:

$$\min_t \max_{\eta\leqslant\lambda\leqslant 1}\left|\frac{t-\lambda}{t+\lambda}\right| = \min \max\left\{\left|\frac{t-\eta}{t+\eta}\right|, \left|\frac{t-1}{t+1}\right|\right\}$$

and t is found from

$$-\frac{t-\eta}{t+\eta} = \frac{t-1}{t+1}, \text{ that is, } t = \sqrt{\eta}.$$

Thus, the computation of the optimal sequence of parameters of length 2^s proceeds as follows:

1. We find η_s, p, q and r by the formulae (7.3) in terms of the spectral bounds l', L', l'', L''.

2. Next, we determine $\eta_{s-1}, \eta_{s-2}, \ldots, \eta_0$ from the recurrence formula

$$\eta_i = \frac{2\sqrt{\eta_{i+1}}}{1+\eta_{i+1}} \quad (i=s-1, s-2, \ldots, 0).$$

3. We put $t_1^{(0)} = \sqrt{\eta_0}$ and consecutively compute 2^{i+1} parameters $t_j^{(i+1)}$ in terms of the 2^i known numbers $t_j^{(i)}$ $(j=1, 2, \ldots, 2^i)$

$$t_j^{(i+1)} = \frac{1+\eta_{i+1}}{2}t_j^{(i)} - \sqrt{\left[\frac{1+\eta_{i+1}}{2}t_j^{(i)}\right]^2-\eta_{i+1}},$$

$$t_{2^i+1-j}^{(i+1)} = \frac{1+\eta_{i+1}}{2}t_j^{(i)} + \sqrt{\left[\frac{1+\eta_{i+1}}{2}t_j^{(i)}\right]^2-\eta_{i+1}} \quad (j=1, 2, \ldots, 2^i).$$

4. Having found the 2^s numbers $t_i^{(s)}$, $i = 1, 2, \ldots, 2^s$, we compute the parameters τ_i' and τ_i'' of the iteration algorithm

$$\tau_i' = \frac{q+rt_i}{t_i+p}, \quad \tau_i'' = \frac{rt_i-q}{p-t_i} \quad (i=1, 2, \ldots, 2^s).$$

The efficiency of the iteration process is easily found. For we have shown above that

$$\min_{\{\tau\}_s}\max_{\eta_s\leqslant\lambda\leqslant 1}\left|\prod_{i=1}^{s}\frac{1-\tau_i\lambda}{1+\tau_i\lambda}\right| = \min_{\{\tau\}_{s-1}}\max_{\eta_{s-1}\leqslant\lambda\leqslant 1}\left|\prod_{i=1}^{2^{s-1}}\frac{1-\tau_i\lambda}{1+\tau_i\lambda}\right| = \ldots$$

$$\ldots = \min_{\tau}\max_{\eta_0\leqslant\lambda\leqslant 1}\left|\frac{1-\tau\lambda}{1+\tau\lambda}\right| = \frac{1-\sqrt{\eta_0}}{1+\sqrt{\eta_0}} = q(\eta, s).$$

In $\nu = 2^s$ iterations (7.1) we obtain an approximation whose residual and error are estimated by

$$\|r^\nu\| \leqslant \|r^0\|q^2(\eta, s), \quad \|v^\nu\| \leqslant \|v^0\|q^2(\eta, s).$$

The average efficiency of the process (7.1) (per iteration) is $\varkappa = -\frac{2}{2^s}\ln q(\eta, s)$. In [8] there is an estimate of $q^2(\eta, s)$, which shows that

$$q^2(\eta, s) \leqslant 4\exp\frac{\pi^2\nu}{\ln(\eta/4)} \quad (\nu = 2^s).$$

We do not reproduce the argument; an idea of the efficiency of the method can be gained from Table 3,

η \ s	1	2	3	4
0.002	3.83	5.92	7.00	7.53
0.001	3.55	5.82	7.01	7.62
0.0005	3.26	5.71	7.02	7.69
0.00025	2.98	5.58	7.00	7.75
0.000125	2.71	5.44	7.00	7.78
0.0000625	2.44	5.29	6.98	7.79

which gives, for $s = 1, 2, 3$ and 4 and various η, the values of

$$\gamma(\eta, s) = -\frac{2}{2^s} \ln q(\eta, s) \cdot \ln \eta = \varkappa \ln \eta^{-1}.$$

It is clear that γ is almost independent of η and that, for $s \geqslant 3$, it does not change very much with s. The decrease of residuals (on average) proceeds according to the formula

$$\| r^\nu \| \simeq \| r^0 \| e^{\nu \frac{\gamma}{\ln \eta}} \quad (\eta < 1!).$$

For Poisson's problem $\Delta u = f$ with the 100 x 100 mesh ($\eta = 0.00025$), a cycle of 8 iterations decreases the residuals approximately 900 times and a cycle of 16 iterations 3×10^6 times.[1]

PROOF OF LEMMA 2. Suppose that the sequence $\{t\}_n = \{t_1, t_2, \ldots, t_n\}$ realizes

$$\min_{\{t\}_n} \max_{\eta \leqslant \lambda \leqslant 1} \left| \prod_{i=1}^{n} \frac{t_i - \lambda}{t_i + \lambda} \right| = q.$$

It is easy to obtain some properties of $\{t\}_n$ ($t_i \leqslant t_{i+1}$).

1) All the t_i lie in $[\eta, 1]$. If $t_1 \leqslant \eta$, then $\frac{t_1 - \lambda}{t_1 + \lambda} \leqslant 0$ on $[\eta, 1]$, $\frac{d}{dt_1} \left| \frac{t_1 - \lambda}{t_1 + \lambda} \right| = -\frac{d}{dt_1} \frac{t_1 - \lambda}{t_1 + \lambda} = -\frac{2\lambda}{(t_1 + \lambda)^2} < 0$, and such a sequence cannot be optimal.

In the same way we show that $t_n < 1$.

2. Between any two zeros t_j and t_{j+1} the function $\left| \prod_{i=1}^{n} \frac{t_i - \lambda}{t_i + \lambda} \right|$ attains its maximum q.

Assume that this is not so and let

$$\max_{t_j \leqslant \lambda \leqslant t_{j+1}} \left| \prod_{i=1}^{n} \frac{t_i - \lambda}{t_i + \lambda} \right| = q' < q,$$

We compute the derivatives along some direction in the (t_j, t_{j+1})-plane

$$\left(\frac{d}{dt_j} - \beta \frac{d}{dt_{j+1}} \right) \left(\frac{t_j - \lambda}{t_j + \lambda} \frac{t_{j+1} - \lambda}{t_{j+1} + \lambda} \right) = 2\lambda \frac{(t_{j+1}^2 - \beta t_j^2) - \lambda^2 (1 - \beta)}{(t_j + \lambda)^2 (t_{j+1} + \lambda)^2}.$$

We choose β so that $t_{j+1}^2 - \beta t_j^2 > 0$ and $\beta > 1$, that is, $1 < \beta < t_{j+1}^2 / t_j^2$.

The function $\frac{t_j - \lambda}{t_j + \lambda} \frac{t_{j+1} - \lambda}{t_{j+1} + \lambda}$ is positive on $[\eta, t_j) \cup (t_{j+1}, 1]$ and negative on (t_j, t_{j+1}). Hence

$$-\left(\frac{d}{dt_j} - \beta \frac{d}{dt_{j+1}} \right) \left| \frac{t_j - \lambda}{t_j + \lambda} \frac{t_{j+1} - \lambda}{t_{j+1} + \lambda} \right| \begin{cases} < 0 \text{ on } [\eta, t_j) \cup (t_{j+1}, 1], \\ > 0 \text{ on } (t_j, t_{j+1}). \end{cases}$$

[1] That is, $\| r^\nu \| \simeq \| r^0 \| \cdot e^{-0.9 \cdot \nu}$.

which shows that $\{t\}_n$ cannot be optimal.

3) The function $\left| \prod_{i=1}^{n} \frac{t_i - \lambda}{t_i + \lambda} \right|$ attains its maximum q at the ends $\lambda = \eta$ and $\lambda = 1$ of the interval.

For $\prod_{i=1}^{n} \frac{t_i - \lambda}{t_i + \lambda}$ is on $[\eta, t_1]$ a product of monotone decreasing positive functions. Hence it is monotone and

$$\max_{\eta \leqslant \lambda \leqslant t_1} \left| \prod_{i=1}^{n} \frac{t_i - \lambda}{t_i + \lambda} \right| = \prod_{i=1}^{n} \frac{t_i - \eta}{t_i + \eta} = q'.$$

Then, as above,

$$\frac{d}{dt_1} \left| \frac{t_1 - \lambda}{t_1 + \lambda} \right| \begin{cases} > 0 \text{ on } [\eta, t_1), \\ < 0 \text{ on } (t_1, 1], \end{cases}$$

and, if $q' < q$ $\{t\}_n$ cannot be optimal.[1]

Thus, we have proved the existence of Chebyshev's alternation, that is, points $\lambda_0 = \eta < \lambda_1 < \lambda_2 < \ldots < \lambda_n = 1$ at which

$$\prod_{i=1}^{n} \frac{t_i - \lambda_j}{t_i + \lambda_j} = (-1)^j q \quad (j = 0, 1, \ldots, n).$$

4) The sequence $\{t_1, t_2, \ldots, t_n\}$ is unique (its existence follows from the continuity in $\{t\}$ of the function $\left| \prod \frac{t_i - \lambda}{t_i + \lambda} \right|$). Assume that there exists another such sequence $\{\tau_1, \tau_2, \ldots, \tau_n\}$. Then the functions $Q(t, \lambda)$ and $Q(\tau, \lambda)$, which have each a Chebyshev alternation on $[\eta, 1]$ coincide in at least n distinct points $\lambda \in [\eta, 1]$. Let $P(\lambda) = Q(t, \lambda) - Q(\tau, \lambda)$:

$$P(\lambda) = \frac{\prod_{i=1}^{n} (t_i - \lambda)(\tau_i + \lambda) - \prod_{i=1}^{n} (\tau_i - \lambda)(t_i + \lambda)}{\prod_{i=1}^{n} (t_i + \lambda)(\tau_i + \lambda)},$$

and the zeros of $P(\lambda)$ must be zeros of the numerator. But the numerator, which is a polynomial of degree $2n - 1$, vanishes at $\lambda = 0$, and is an odd function of λ. As it has n more zeros on $[\eta, 1]$, it must be identically zero.

We consider now the function ($t = \{t\}_n$, $\tau = \{\tau\}_n$)

$$Q(t, \lambda, \tau, \mu) \equiv \prod_{i=1}^{n} \frac{\tau_i - \lambda}{t_i + \lambda} \frac{t_i - \mu}{\tau_i + \mu}$$

[1] A similar argument applies to the right end-point $\lambda = 1$ of the interval.

and establish similar properties for the sequence $\{t, \tau\}_n$ which realizes

$$\min_{\{t,\tau\}n} \max_{\eta\leqslant\lambda,\,\mu\leqslant 1} |Q(t, \lambda, \tau, \mu)| = q.$$

1) All the t_i and $\tau_i \in [\eta, 1]$. This follows from the fact that

$$\frac{d}{dt_i}\frac{t_i-\mu}{t_i+\lambda} = \frac{\lambda+\mu}{(t_i+\lambda)^2} > 0 \text{ and } \eta\leqslant\lambda,\ \mu\leqslant 1$$

and that a sequence $\{t, \tau\}$ with, say, $t_1 \leqslant \eta$ cannot be optimal, because then $\frac{d}{dt_1}\left|\frac{t_1-\mu}{t_1+\lambda}\right| < 0$ for $\eta \leqslant \lambda, \mu \leqslant 1$.

2) The function $|Q(t, \lambda, \tau, \mu)|$ attains its maximum q in every strip $(t_j \leqslant \mu \leqslant t_{j+1}) \times (\mu \leqslant \lambda \leqslant 1)$.

We compute

$$\left(\frac{d}{dt_{j+1}} - \beta\frac{d}{dt_j}\right)\left(\frac{t_j-\mu}{t_j+\lambda}\frac{t_{j+1}-\mu}{t_{j+1}+\lambda}\right) = \frac{(\lambda+\mu)\,p(\lambda, \mu)}{(t_j+\lambda)^2(t_{j+1}+\lambda)^2},$$

where $p(\lambda, \mu) = (t_j^2 - \beta t_{j+1}^2) + (\lambda - \mu)(t_j - \beta t_{j+1}) - (1 - \beta)\lambda\mu$. To prove this it is sufficient to choose β such that $p(\lambda, \mu)$ does not change sign in the square in question, that is, the curve $p(\lambda, \mu) = 0$ does not intersect this square. By putting $a = t_j^2 - \beta t_{j+1}^2$, $b = t_j - \beta t_{j+1}$, $c = 1 - \beta$ we obtain for the curve $p(\lambda, \mu) = 0$ the explicit formula

$$\mu_0(\lambda) = \frac{a+b\lambda}{b+c\lambda}.$$

The behaviour of this curve is such that, if $p(\lambda, \mu)$ does not change its sign on the vertices of the square $[0, 1] \times [0, 1]$, then $p(\lambda, \mu)$ does not change its sign on the whole square. Thus, to determine β we have the four inequalities:

1. $p(0, 0) < 0$, that is, $\beta \geqslant \left(\frac{t_j}{t_{j+1}}\right)^2$,
2. $p(1, 1) < 0$, that is, $\beta < \frac{1-t_j^2}{1-t_{j+1}^2}$,
3. $p(1, 0) < 0$, that is, $\beta \leqslant \frac{t_j(1-t_j)}{t_{j+1}(1-t_{j+1})}$,
4. $p(0, 1) < 0$, that is, $\beta \geqslant \frac{t_j(1+t_j)}{t_{j+1}(1+t_{j+1})}$.

It is easily checked that these inequalities are compatible for any $0 < t_j < t_{j+1}$. Hence there exists a β such that the derivative

$\left(\frac{d}{dt_{j+1}} - \beta\frac{d}{dt_j}\right)\left|\frac{t_j-\mu}{t_j+\lambda}\frac{t_{j+1}-\mu}{t_{j+1}+\lambda}\right|$ has one sign in the strip $t_j < \mu < t_{j+1}$ and the opposite sign in the rest of the square. The assertion 2) is now proved.

Clearly, the same applies to the bands $\tau_i < \lambda < \tau_{i+1}$.

3) The function $|Q(t, \lambda, \tau, \mu)|$ attains its maximum value q on the boundary of the square. Suppose that it is not so and let

$\max_{\eta\leqslant\mu\leqslant 1} |Q(t, \eta, \tau, \mu)| = q' < q$. However, $\prod_{i=1}^{n} \frac{\tau_i-\lambda}{t_i+\lambda}$ is monotonic in the strip $\eta \leqslant \lambda \leqslant \tau_1$, (because it is a product of positive monotonic functions). Hence

$$\max_{\eta\leqslant\lambda\leqslant\tau_1} \max_{\eta\leqslant\mu\leqslant 1} |Q(t, \lambda, \tau, \mu)| = \max_{\eta\leqslant\mu\leqslant 1} |Q(t, \eta, \tau, \mu)| = q' < q.$$

But $\frac{d}{d\tau_1}\left|\frac{\tau_1-\lambda}{\tau_1+\mu}\right|$ is positive in the strip $\eta \leqslant \lambda \leqslant \tau$ and negative in the remaining part of the square. And in that case the assumption $q' < q$ is incompatible with the optimality of $\{t, \tau\}_n$.

4) In the optimal sequence $\{t, \tau\}_n$ we have $t_i = \tau_i$ for $i = 1, 2, \ldots, m$ and consequently the solution of the minimax problem is unique.

Put $Q(t, \tau, \lambda) \equiv \prod_{i=1}^{n} \frac{\tau_i-\lambda}{t_i+\lambda}$. Then $Q(t, \lambda, \tau, \mu) = Q(t, \tau, \lambda)Q(\tau, t, \mu)$ and

$$\max_{\lambda,\,\mu} |Q(t, \lambda, \tau, \mu)| = \max_{\lambda} |Q(t, \tau, \lambda)| \max_{\mu} |Q(\tau, t, \mu)|.$$

Let $q_1 = \max_{\lambda} |Q(t, \tau, \lambda)|$, $q_2 = \max_{\mu} |Q(\tau, t, \mu)|$. Let $\lambda_k(\mu_p)$ be the maximum point of $|Q(t, \tau, \lambda)|$ (or $|Q(\tau, t, \mu)|$) on $[\eta, 1]$. Then $q = q_1 q_2$, and the set of maximum points of $|Q(t, \lambda, \tau, \mu)|$ forms a rectangular mesh $\{\lambda_k, \mu_p\}_{k,p}$ in the square. It has been shown above that $|Q(t, \lambda, \tau, \mu)|$ attains its maximum on the boundary of the square and in the strips $\tau_i \leqslant \lambda \leqslant \tau_{i+1}$, $t_j < \mu < t_{j+1}$. Hence the points λ_k (μ_p) form a Chebyshev's alternation for $Q(t, \tau, \mu)$ (or $Q(\tau, t, \mu)$) which contains the points $\lambda = \eta$ and $\lambda = 1$. But the functions $Q(t, \tau, \lambda)$ and $Q(\tau, t, \lambda)$ which both admit an alternation, although possibly with different amplitudes q_1 and q_2, must coincide at at least n points of the interval $[\eta, 1]$. Hence, there are at least n zeros on $[\eta, 1]$ of

$$P(\lambda) \equiv \prod_{i=1}^{n} (\tau_i-\lambda)(\tau_i+\lambda) - \prod_{i=1}^{n} (t_i-\lambda)(t_i+\lambda).$$

The polynomial $P(\lambda)$ is even of degree $2n - 2$, and so it can have n positive zeros only if $P(\lambda) \equiv 0$, that is $\tau_i = t_i$.

§ 8. Further development of the method of alternating directions

The theory of choosing highly efficient iteration parameters for the method of alternating directions rests essentially on two factors: the system of difference equations is given in the form

$$D_1 u + D_2 u = f, \tag{8.1}$$

where

1) D_1 and D_2 are commuting negative-definite self-adjoint operators (on the finite-dimensional space of the mesh functions);

2) the solution of the equations $(E - \tau D_i)v = \varphi$ $(i = 1, 2)$ is much simpler than that of the original problem.

From the point of view of content this leads to the following circle of problems:

1. The domain is rectangular (to be specific, the $(\pi \times \pi)$-square); in other domains the variables do not separate and the operators D_1 and D_2 do not commute.[1]

2. The most general form of a self-adjoint equation with separating variables (commutativity!) is

$$(8.2)\qquad \begin{cases} \frac{\partial}{\partial x} a(x) \frac{\partial u}{\partial x} + \frac{\partial}{\partial y} b(y) \frac{\partial u}{\partial y} + c_1(x) u + c_2(y) u = f, \\ D_1 u \equiv \frac{\partial}{\partial x} a(x) \frac{\partial u}{\partial x} + c_1(x) u, \\ D_2 u \equiv \frac{\partial}{\partial y} b(y) \frac{\partial u}{\partial y} + c_2(y) u. \end{cases}$$

3. There are general boundary conditions of the third kind, $\alpha \frac{\partial u}{\partial x} + \beta u' = \varphi$, but α and β must be constant on every edge of the square.

In that case we consider the following operators acting on functions of one variable:

$$(8.3)\qquad \begin{cases} D_1 u = \frac{\partial}{\partial x} a \frac{\partial u}{\partial x} + c_1 u, \quad \alpha_1 \frac{\partial u}{\partial x} + \beta u|_{x=0} = 0, \\ \qquad\qquad \alpha_2 \frac{\partial u}{\partial x} + \beta_2 u|_{x=\pi} = 0, \\ D_2 u \equiv \frac{\partial}{\partial y} b \frac{\partial u}{\partial y} + c_2 u, \quad \alpha_3 \frac{\partial u}{\partial y} + \beta_3 u|_{y=0} = 0, \\ \qquad\qquad \alpha_4 \frac{\partial u}{\partial y} + \beta_4 u|_{y=\pi} = 0. \end{cases}$$

These operators are self-adjoint; they are negative-definite under certain restrictions on the coefficients a, b, α and β, which we do not intend to analyse. As operators acting on functions $u(x, y)$ defined on the $(\pi \times \pi)$-square and satisfying the homogeneous boundary conditions, D_1 and D_2 commute.[2]

[1] More precisely, one does not succeed in choosing D_1 and D_2 so that $\Delta u = D_1 u + D_2 u$ and the conditions 1) and 2) are satisfied.

[2] The computational scheme of the method of alternating directions makes fairly strong use of the fact that the differential operator is approximated by the simplest five-point scheme. The solution of difference equations of a higher order of accuracy (say, $O(h^4)$ or $O(h^6)$) requires some modifications even for the simplest equation $\Delta u = f$. These modifications were proposed by Samarskii ([11], 466) in both the two-dimensional and three-dimensional cases. The analysis of convergence leads to functions on the spectrum similar to those considered above, and the choice of the sequence of iteration parameters can be based on an approximate as well as on the exact solution of the minimax problem.

Let us now consider the class of concrete problems to which the algorithmic scheme of the method of alternating directions is applicable, though at present without a theoretical guarantee of success.

1. The form of the equation can be quite general, the only restriction being the absence of the mixed derivative

$$(8.4)\qquad \frac{\partial}{\partial x} a(x, y) \frac{\partial u}{\partial x} + \frac{\partial}{\partial y} b(x, y) \frac{\partial u}{\partial y} + A(x, y) \frac{\partial u}{\partial x} + B(x, y) \frac{\partial u}{\partial y} + C(x, y) u = f.$$

2. The domain is practically arbitrary if the first boundary-value problem is to be solved.

If the boundary consists of straight-line segments parallel to the coordinate axes, boundary conditions of the third kind are permissible: $\alpha \frac{\partial u}{\partial n} + \beta u = \varphi$, with variable α and β. The operators D_1 and D_2 are then defined in the obvious manner. For example,

$$(D_1 u)_{n,m} = \left(\frac{\partial}{\partial x} a \frac{\partial u}{\partial x}\right)_{n,m} + A_{n,m} \frac{u_{n+1,m} - u_{n-1,m}}{2h} + C_{n,m} \cdot u_{n,m},$$

and the system of difference equations

$$u_{n,m} - \frac{\tau}{h^2}\left[a_{n+\frac{1}{2},m}(u_{n+1,m} - u_{n,m}) - a_{n-\frac{1}{2},m}(u_{n,m} - u_{n-1,m})\right] - \frac{\tau}{2h} A_{n,m}(u_{n+1,m} - u_{n-1,m}) + \tau C_{n,m} \cdot u_{n,m} = r_{n,m},$$

with boundary conditions of the first kind on an arbitrary domain, or boundary conditions of the third kind if $\frac{\partial u}{\partial n}$ is $\frac{\partial u}{\partial x}$ or $\frac{\partial u}{\partial y}$ is easily solved by sweeping.

A numerical experiment has shown that with the formal application of the algorithmic scheme of the method of alternating directions the convergence of the iterations often remains high even if the existing theory is not applicable. For example, a large number of computations for the Laplace equation has been performed by various authors in a domain obtained by deleting parts of the $(\pi \times \pi)$-square. As an example, consider the first boundary-value problem for $\Delta u = 0$ in the domain shown in Fig. 6. The iterations proceed according to (6.1) with the choice of the iteration parameters given by the approximate theory for the containing $(\pi \times \pi)$-square (the mesh step $h = \pi/100$: $\tau_1 = 0.00056$, $\tau_2 = 0.00296$, $\tau_3 = 0.0156$, $\tau_4 = 0.0825$, $\tau_5 = 0.436$. The decrease of the residual $\| \Delta u \|$ is shown in Table 4.

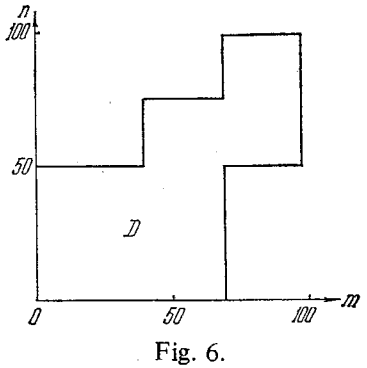


Fig. 6.

The initial approximation was taken to be zero inside the domain, with the prescribed values on the boundary. The convergence of the iterations stops after the 25^{th} iteration; this is due to the influence of the rounding error $\left(\Delta u^{\nu} \simeq \frac{u\varepsilon}{h^2}\right)$. A similar pattern was observed in the solution of problems in other domains of similar shape: the computation parameters for the containing square give convergence close to the theoretical values (for the square) (see [9] and [10]). However, attempts to generalize the theoretical analysis lead to more modest results than the experiments.

ν	0	5	10	15	20	25	30	35	40
$\|\Delta u\|$	91.15	.85	.064	.0047	.00041	.00026	.00017	.00026	.00018

The convergence of the method of alternating directions for non-commuting operators. We first mention the papers in which the analysis of convergence of the method of alternating directions is not based on the assumption that D_1 and D_2 in (8.1) commute. In [9] and [11] convergence of the iterations for (8.1) is studied under the sole assumption that D_1 and D_2 are negative-definite self-adjoint operators, without assuming that they commute. The corresponding concrete class of problems is nearly the same as the class of problems that allow a formal application of the algorithmic scheme (nearly, because self-adjointness breaks down for the terms $A\frac{\partial u}{\partial x}$ and $B\frac{\partial u}{\partial y}$). However, the results obtained in this case are relatively modest: suppose, as before, that the spectra of $-D_1$ and $-D_2$ lie in the intervals $[l', L']$ and $[l'', L'']$, where l' and $l'' > 0$. Then the iterations with the fixed parameter $\tau = 1/\sqrt{lL}$, $l = \min(l', l'')$ and $L = \max(L', L'')$,

$$(E - \tau D_1)u^* = (E + \tau D_2)u^{\nu} - \tau f,$$
$$(E - \tau D_2)u^{\nu+1} = (E + \tau D_1)u^* - \tau f,$$

converge, and the rate of convergence of the error in the iteration process is given by the formula

$$\|v^{\nu}\| \simeq \|v^0\| e^{-\nu 2\sqrt{\frac{l}{L}}} \tag{8.5}$$

This estimate (for a single parameter τ) cannot be improved; it is easy to check in the example of the equation $\Delta u = f$ that the bound (8.5) is attained on the square. In this form, if we disregard the unexploited possibility of varying the parameter τ_ν, this method has no advantage over Richardson's method in the rate of convergence, and it is substantially inferior to it in its scope: Richardson's method can be applied to problems with mixed derivatives.

Significant progress in the analysis of the case when the operators do not commute was achieved in [12], however, at the expense of considerably narrowing the class of problems. By considering a finite-difference approximation to the equation

$$\frac{\partial}{\partial x}a(x, y)\frac{\partial u}{\partial x} + \frac{\partial}{\partial y}b(x, y)\frac{\partial u}{\partial y} = f(x, y) \tag{8.6}$$

(where the domain is a square, the boundary conditions are of the first kind, and the five-point difference scheme is used) for sufficiently smooth coefficients $a(x, y)$ and $b(x, y)$, Widlund constructs the iteration process

$$\begin{aligned}\left(\tau'_{\nu}\Lambda - \frac{\partial}{\partial x}a\frac{\partial}{\partial x}\right)u^* &= \left(\tau'_{\nu}\Lambda + \frac{\partial}{\partial y}b\frac{\partial}{\partial y}\right)u^{\nu} - f,\\ \left(\tau''_{\nu}\Lambda - \frac{\partial}{\partial y}b\frac{\partial}{\partial y}\right)u^{\nu+1} &= \left(\tau''_{\nu}\Lambda + \frac{\partial}{\partial x}a\frac{\partial}{\partial x}\right)u^* - f,\end{aligned} \tag{8.7}$$

where Λ is a diagonal matrix whose entries are computed in a particular way from $a(x, y)$ and $b(x, y)$. He proposes a method for choosing a sequence of ν pairs of iteration parameters $\{\tau'_i, \tau''_i\}_{i=1}^{\nu}$ and obtains the estimate

$$\|v^{\nu}\| \leqslant \|v^0\|\left[1 - c\left(\frac{1}{N}\right)^{1/\nu}\right]. \tag{8.8}$$

Thus, the average efficiency of this process is

$$\varkappa = -\frac{1}{\nu}\ln\left[1 - c\left(\frac{1}{N}\right)^{1/\nu}\right] \simeq O\left(\frac{1}{\nu}\frac{1}{N^{1/\nu}}\right). \tag{8.9}$$

The optimal length ν of the iteration cycle can be computed from (8.9). We refrain from discussing the practical applications of this result in detail, because in this situation, and for sufficiently smooth $a(x, y)$ and $b(x, y)$, it is evidently preferable to use the ordinary method of alternating directions with parameters τ_i computed for the equation $au_{xx} + bu_{yy} = f$, where a and b are the mean values of the coefficients $a(x, y)$ and $b(x, y)$ over the domain. However, the modification (8.7) of the method of alternating directions may prove useful exactly in the case of strongly oscillating $a(x, y)$ and $b(x, y)$. (The main element of this modification is the replacement of the identity matrix by the special diagonal matrix Λ).

In the estimates (8.5) and (8.8) there is a certain lack of rigour; the fact of the matter is that these results are established not for the ordinary norm of the mesh function; they are valid in terms of specially constructed norms, which are close to the so-called integral energy. However, this circumstance does not influence the value of these methods as a practical tool for the solution of finite-difference elliptic equations. But we do not dwell on this in detail.

The computational scheme of D'yakonov [13]. One general idea of constructing iteration processes for the solution of the equation $Du = f$ has

been known for a long time. Namely it is recommended to use the process[1]

$$(8.10)\qquad B\frac{u^{\nu+1}-u^{\nu}}{\tau}=Du^{\nu}-f \quad \text{or} \quad u^{\nu+1}=(E+\tau B^{-1}D)\,u^{\nu}-\tau B^{-1}f.$$

The properties of B that are necessary to ensure the efficiency of the iterations (8.10) are also known:

1. B must be easily invertible (in the sense that the determination of u from the equation $Bu = z$ for u must be considerably simpler than the original problem).

2. The norm of the operator $(E + \tau B^{-1}D)$ must be as small as possible, that is, $E + \tau B^{-1}D \simeq 0$, which means that B is in a certain sense close to D. Unfortunately, it is so difficult to satisfy these requirements together that the idea on its own, without a suitable construction for B, is of little value.

D'yakonov apparently was one of the first people who drew attention to the possibilities of combining this general construction with the ideas and achievements of the method of alternating directions. He proposed several constructions of B and developed the corresponding theory. If we disregard the secondary details, we can distinguish at present three constructions of B.

I. The finite-difference analogue of the Laplace operator, $B \equiv \Delta$.

II. The finite-difference analogue of the fourth order operator.

$$(8.11)\qquad B\equiv\left(E-\sigma_1\frac{\partial^2}{\partial x^2}\right)\left(E-\sigma_2\frac{\partial^2}{\partial y^2}\right).$$

The degeneracy of this operator makes itself felt in the fact that the boundary conditions sufficient for the existence of B^{-1} are the same as for a second order elliptic operator.

III. The finite-difference analogue of the degenerate elliptic equation

$$(8.12)\qquad B\equiv\left\{E-\sigma\left(\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)\right\}\cdot\left\{E+\sigma\left(\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)\right\}\equiv E-\sigma^2\left(\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)^2.$$

Regarded as a differential operator, (8.12) can be equally well be taken to be a degenerate hyperbolic operator. However, the presence of E in the finite-difference approximation gives to (8.12) for sufficiently small values of σ typically elliptic features.

The constructions (8.11) and (8.12) contain the parameters σ_1, σ_2 and σ and the theory should recommend the optimal choices (in some sense) of

[1] Note that the operator $B^{-1}D$ is not self-adjoint if B and D do not commute, but it can be replaced by a self-adjoint operator after a change of variables. For the iteration $u^{\nu+1}=u^{\nu}+\tau B^{-1}Du^{\nu}-\tau B^{-1}f$ after multiplying by $D^{1/2}$ and setting $\tilde{u}=D^{1/2}u$ turns into $\tilde{u}^{\nu+1}=\tilde{u}^{\nu}+\tau(D^{1/2}B^{-1}D^{1/2})\,u^{\nu}-\tau\tilde{f}$, $\tilde{f}=D^{1/2}B^{-1}f$, where $D^{1/2}B^{-1}D^{1/2}$ is a self-adjoint operator. Convergence in terms of u and v is studied in the usual way and leads to estimates for the norm of v. Returning to the original function we note that $\|\tilde{v}\|=(\tilde{v},\tilde{v})^{1/2}=(D^{1/2}v, D^{1/2}v)=(Dv, v)^{1/2}$, so that the matter reduces to a different definition of the norm of the error v. (The operators B and D are assumed to be self-adjoint, and D is assumed to be positive.)

these parameters.[1] To invert any of the three operators B we need some boundary conditions, and the boundary conditions[2] of the original problem are mostly used. However, in principle we may also use a different method, in which B is supplemented by some simpler boundary conditions; then the function

$$u^* = (E + \tau B^{-1}D)u^{\nu} - \tau B^{-1}f$$

does not satisfy the boundary conditions of the original problem and its values at the boundary mesh points have to be readjusted so as to turn the function u^* into $u^{\nu+1}$. This procedure complicates the question of convergence, but sometimes one has to resort to it when the original boundary conditions are too complicated for B to be easily invertible. An example of the use of this scheme in a problem from elasticity theory with complicated non-local boundary conditions can be found in [14].

Let us consider in detail questions arising in the practical application of these constructions.

I. $B \equiv \Delta$ (a good choice from the point of view of the second requirement imposed on B). It is standard knowledge in the theory of elliptic equations that given a strongly[3] elliptic operator D there exist constants c, C and C' such that for all u

$$(8.13)\qquad c(-\Delta u, u)\leqslant (Du, u)\leqslant C(-\Delta u, u)\leqslant C'(Du, u).$$

For our purposes it is essential that as a consequence of (8.13) the operator $-D^{1/2}\Delta^{-1}D^{1/2}$ is bounded below and above:

$$(8.14)\qquad l(u, u)\leqslant -(D^{1/2}\Delta^{-1}D^{1/2}u, u)\leqslant L(u, u),\quad l>0.$$

Thus, the spectrum of $-D^{1/2}\Delta^{-1}D^{1/2}$ lies in $[l, L]$. It is natural to expect a similar inequality to (8.14) for the finite-difference approximations of D and Δ, with l and L practically independent of the mesh size. Combining (8.10) with Richardson's method

$$\Delta\frac{u^{\nu+1}-u^{\nu}}{\tau_{\nu+1}}=Du^{\nu}-f,$$

we obtain a process converging on average according to the formula:

$$(8.15)\qquad \|v^{\nu}\|\simeq\|v^{0}\|\,e^{-\nu 2\sqrt{(l/L)}}.$$

However, each iteration with $B \equiv \Delta$ requires the solution of the equation $\Delta u = \varphi$, and we only have iteration methods to do that.[4] This fact

[1] In this scheme the parameters are usually optimized separately: σ is chosen so that the ratio of the spectral bounds for $D^{1/2}B^{-1}D^{1/2}$ is as close to 1 as possible, and then the τ_ν are chosen, for example, by the same arguments as in Richardson's method.

[2] Homogeneous!

[3] That is, there exists a constant $\delta > 0$ such that $(Du, u)\geqslant\delta(u, u)$.

[4] It is worth drawing attention to the 'rapid Fourier transform' [15]. In its trivial form the Fourier transform method, which requires the computation of $\sim N^2$ integrals over N^2 points, does not stand up to the competition of the iteration methods either in the number of operations (which is $O(N^4)$) or in the possibility of generalizations: it is applicable virtually only to the Poisson equation $\Delta u = f$ on a rectangle. However, Hockney succeeds in constructing an extremely rational scheme of computations, which requires only $O([N\ln N]^2)$ operations (for $N = 2^p$).

determines the range of efficiency of the whole process: the operator D can be quite general, but the domain and the boundary conditions must be such that the method of alternating directions with a good choice of parameters (relating, so to speak, to the intrinsic iteration cycle for the solution of the equation $\Delta u = \varphi$) is about as efficient as on the rectangle: one must not forget about the competition from Richardson's method ($B \equiv E$). In a rectangular domain the number of iterations needed to decrease the residuals ε^{-1} times is of the order

$$O\left(\ln N \cdot \ln \frac{1}{\varepsilon_1} \cdot \ln \frac{1}{\varepsilon} \cdot \sqrt{(L/l)}\right).$$

Here $O\left(\ln N \cdot \ln \frac{1}{\varepsilon_1}\right)$ is the number of inner iterations for solving (with the accuracy ε_1) $\Delta u = \varphi$ and $\ln \frac{1}{\varepsilon} \cdot \sqrt{(L/l)}$ is the number of basic iterations (8.10).

The practical use of the method requires the solution of two more problems.

1) An estimate of the bounds l and L of the spectrum of $D^{1/2}\Delta^{-1}D^{1/2}$. We note that for this operator the smallest eigenvalue no longer corresponds to the smoothest eigenfunction; this can be checked in the simplest example, by taking

$$-D \equiv \frac{\partial^2}{\partial x^2} + a\frac{\partial^2}{\partial y^2}, \qquad 0 \leqslant x,\ y \leqslant \pi.$$

2) It is not, apparently, necessary to insist on high accuracy in solving the equation $\Delta u = \varphi$. The question arises of deciding the optimal number of 'inner' iterations, related possibly to the ordinal ν of the basic iteration.[1]

D'yakonov considers the process with fixed τ and the 'inversion' of Δ is effected in a single cycle of iterations of the method of alternating directions. Thus, the computation proceeds as follows:

1. We find the residual $f^\nu = Du^\nu - f$.
2. This is followed by one cycle (of length $O(\ln N)$) of iterations by the method of alternating directions for solving the equation $\Delta v = r^\nu$ (beginning with the approximation $v^0 \equiv 0$).
3. The resulting rough approximation $\tilde{v}$ is used to correct the values of u: $u^{\nu+1} = u^\nu + \tau\tilde{v}$. (Homogeneous boundary conditions are taken for v.)

Gunn [16] suggests a somewhat different scheme: we write (7.1) as $\Delta u^{\nu+1} = \Delta u^\nu + \tau(Du^\nu - f)$ and proceed as follows:

1. $r^\nu = \Delta u^\nu + \tau(Du^\nu - f)$.
2. $u^{\nu+1}$ is found as a rough solution of the equation $\Delta u^{\nu+1} = r^\nu$ (using a single cycle of iterations of the method of alternating directions and the initial approximation u^ν).

In both cases recommendations are given for the choice of τ and estimate the number of iterations as $O\left(\ln N \cdot \ln \frac{1}{\varepsilon}\right)$: $\tau \sim 1/L$.

[1] Nikolaev has recently published a paper on this topic in Zh Vychisl. Mat. i Mat. Fiz 12:6 (1972).

II. $B \equiv \left(E - \sigma_1\frac{\partial^2}{\partial x^2}\right)\left(E - \sigma_2\frac{\partial^2}{\partial y^2}\right)$. In this case the solution of the equation $Bu = \varphi$ is easily obtained by sweeps in the directions of x and y (in an arbitrary domain for the first boundary-value problem, and in a domain bounded by segments parallel to coordinate axes for the third boundary-value problem). As for the second requirement that $D^{1/2}B^{-1}D^{1/2}$ should be close to E, the position is worse than for $B \equiv \Delta$, The possibilities of approximating $\frac{1}{\tau}E$ by $B^{-1}D$ by a choice of the parameters σ_1 and σ_2 can be assessed in the simplest example $D \equiv -\Delta$ on the $(\pi \times \pi)$-square with boundary conditions of the first kind. We denote by λ' and λ'' the eigenvalues of $-\frac{\partial^2}{\partial x^2}$ and $-\frac{\partial^2}{\partial y^2}$ and compute the eigenvalues μ of $B^{-1}\Delta$: $\mu = \frac{\lambda' + \lambda''}{(1+\sigma\lambda')\,(1+\sigma\lambda'')}$. We omit the simple analysis of the spectrum $\mu(\lambda', \lambda'')$; the choice of σ is aimed at making the ratio $\mu_{\min}/\mu_{\max}$ as close to 1 as possible. The optimal values of σ in this sense are $\sigma_1 = \sigma_2 = 1/\sqrt{Ll}$;

$$\mu_{\max} = \mu\left(\frac{1}{\sigma}, \frac{1}{\sigma}\right) = \frac{1}{2\sigma}, \qquad \mu_{\min} = \mu(l, l) = \mu(L, L) \simeq 2l$$

(we are interested in the case $l \ll L$). Thus $\mu_{\min}/\mu_{\max} = 4\sqrt{(l/L)}$. Using the process (8.10) with τ_ν chosen as in Richardson's method, we obtain the decrease in the norm of the error

$$\| v^\nu \| \simeq \| v^0 \| e^{-\nu 4\left(\frac{l}{L}\right)^{1/4}} \simeq \| v^0 \| e^{-\nu O(N^{-1/2})}.$$

III. $B \equiv \left\{E + \sigma\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)\right\}\left\{E - \sigma\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)\right\} \equiv (E + \sigma B_1)(E + \sigma B_2)$. First of all we consider the difference approximations of B_1 and B_2 in a rectangular domain and the algorithm for solving equations of the type

$$(8.16) \qquad \{(E + \sigma B_1)u\}_{n,m} \equiv u_{n,m} + \frac{\sigma}{h}(u_{n,m} - u_{n-1,m}) + \frac{\sigma}{h}(u_{n,m} - u_{n,m-1}) = \varphi_{n,m}.$$

The stencil of the operator is shown in Fig. 7.

Considering for simplicity only the first boundary-value problem (the boundary conditions of the third kind lead to obvious and insignificant changes in computations), we easily determine the unknowns $u_{n,m}$ ($n, m = 1, 2, \ldots, N-1$) from the corresponding equations (8.16)

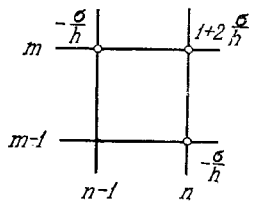


Fig. 7.

starting in the bottom left-hand corner and moving along the rows or columns. In either case, the value of $u_{n-1,\,m}$ and $u_{n,\,m-1}$ are already known at the time of computation of $u_{n,m}$. In exactly the same way we solve the system of equations

$$(8.17)\qquad \{(E+\sigma B_2)\,u\}_{n,\,m} \equiv$$
$$\equiv u_{n,\,m}-\frac{\sigma}{h}\,(u_{n+1,\,m}-u_{n,\,m})-\frac{\sigma}{h}\,(u_{n,\,m+1}-u_{n,\,m})=\varphi_{n,\,m}.$$

The only difference is that we have to start in the top right-hand corner $(N-1,\ N-1)$. The operators B_1 and B_2 have the following obvious properties:

$$(8.18)\quad 1)\ B_1^*=B_2,\quad 2)\ \frac{1}{h}\,(B_1+B_2)=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2},\quad 3)\ B^*=B.$$

This enables us to construct the following iteration process for solving the Poisson equation $\Delta f = f$:

$$(8.19)\qquad \begin{cases} \left(E+\frac{\sigma}{h}B_1\right)u^*=\left(E-\frac{\sigma}{h}B_2\right)u^\nu+\sigma f,\\ \left(E+\frac{\sigma}{h}B_2\right)u^{\nu+1}=\left(E-\frac{\sigma}{h}B_1\right)u^*+\sigma f, \end{cases}$$

For the solution of the equation $[a(x,\,y)u_x]_x + [b(x,\,y)u_y]_y = f$, the operators B_1 and B_2 must be modified in the obvious way, for example

$$(8.20)\qquad (Bu)_{n,\,m}\equiv a_{n-\frac{1}{2},\,m}\,\frac{1}{h}\,(u_{n,\,m}-u_{n-1,\,m})+\frac{b_{n,\,m-\frac{1}{2}}}{h}\,(u_{n,\,m}-u_{n,\,m-1}).$$

The evolution of the error in the process (8.19) is given by

$$(8.21)\qquad v^{\nu+1}=\left(E+\frac{\sigma}{h}B_2\right)^{-1}\left(E-\frac{\sigma}{h}B_1\right)\left(E+\frac{\sigma}{h}B_1\right)^{-1}\left(E-\frac{\sigma}{h}B_2\right)v^\nu\equiv$$
$$\equiv S(\sigma)\,v^\nu.$$

The theory of the optimal choice of the parameter σ which secures $\min \|S(\sigma)\|$ is given in [11] for the case of non-commuting operators B_1^σ and B_2 of the type (8.18). Leaving aside some details, the result of [11] is that, for the optimal choice of σ (which requires a knowledge of the spectral bounds for B_1 and B_2), the process (8.21) converges at the rate

$$\|v^\nu\|\simeq\|v^0\|\,e^{-\nu\cdot 0\sqrt{(l/L)}}$$

(where l and L are the spectral bounds for the operator $(au_x)_x + (bu_y)_y$). Thus, (8.21) is no more efficient and is much more limited in scope than Richardson's method.

However, the easily invertible operator B can also be used in the standard scheme

$$(8.22)\qquad B(\sigma)\,\frac{u^{\nu+1}-u^\nu}{\tau_{\nu+1}}=Du^\nu-f.$$

In this case the parameter σ is chosen so as to make the ratio of the minimal and maximal eigenvalues of $D^{1/2}B^{-1}(\sigma)D^{1/2}$ as close to 1 as possible, and the parameters τ_ν are chosen as in Richardson's method. Corresponding recommendations (for $D\equiv -\Delta$) are given in [10], where (8.22) is called the 'alternating triangular method with Chebyshev acceleration'; in the same paper there is also an estimate of the rate of convergence:

$$(8.23)\qquad \|v^\nu\|\simeq\|v^0\|\,e^{-\nu\cdot 2\sqrt{2}\cdot\sqrt[4]{l/L}}\simeq\|v^0\|\,e^{-\nu O(N^{-1/2})}$$

(where l and L are the spectral bounds of $-\Delta$). In [10] there are also the results of a numerical solution of the equation $\Delta u = f$ in the square, triangle, and ring for various iterative processes. (We remark that Richardson's method in [10] is applied without taking into account the recommendations [1] on ordering the parameters τ_i, and so the degrees of the Chebyshev polynomials are low, of the order 10.)

§ 9. The relaxation method

In [17] the present author proposed the so-called relaxation method for solving difference elliptic equations. In [18] he obtained an estimate of the rate of convergence (for the simplest case of the finite-difference approximation of the Poisson equation $\Delta u = f$ in the square, with boundary conditions of the first kind). It turned out that

$$(9.1)\qquad \|r^\nu\|\simeq\|r^0\|\,e^{-\varkappa\nu}$$

where the convergence exponent $\varkappa$ is independent of the number of mesh-points. In Bakhvalov's paper [19], the convergence of this method is studied in the case of the first boundary-value problem in a rectangle for the general elliptic equation

$$(9.2)\qquad a_{11}\frac{\partial^2 u}{\partial x^2}+2a_{12}\frac{\partial^2 u}{\partial x\,\partial y}+a_{22}\frac{\partial^2 y}{\partial y^2}+a_1\frac{\partial u}{\partial x}+a_2\frac{\partial u}{\partial y}+au=f$$

(where a_{ij}, a_i and a are smooth functions of x and y). For the rate of convergence the same estimate (9.1) is obtained. Finally, Astrakhantsev has considered the convergence of the method for the difference approximation of the third boundary-value problem in an arbitrary two-dimensional domain with smooth boundary for the general self-adjoint elliptic equation

$$(9.3)\qquad \frac{\partial}{\partial x}\,a(x,\,y)\,\frac{\partial u}{\partial y}+\frac{\partial}{\partial y}\,b(x,\,y)\,\frac{\partial u}{\partial x}+\frac{\partial}{\partial x}\,b(x,\,y)\,\frac{\partial u}{\partial y}+$$
$$+\frac{\partial}{\partial y}\,c(x,\,y)\,\frac{\partial u}{\partial y}+Au=f.$$

Using variational difference schemes that guarantee self-adjointness of the finite-difference approximations to (9.3) and assuming that the coefficients a, b, c and A are smooth, he obtains in [20] an estimate of the rate of convergence in a form that differs from (9.1) only in the definition of the

norm. The norm of the error decreases according to the formula

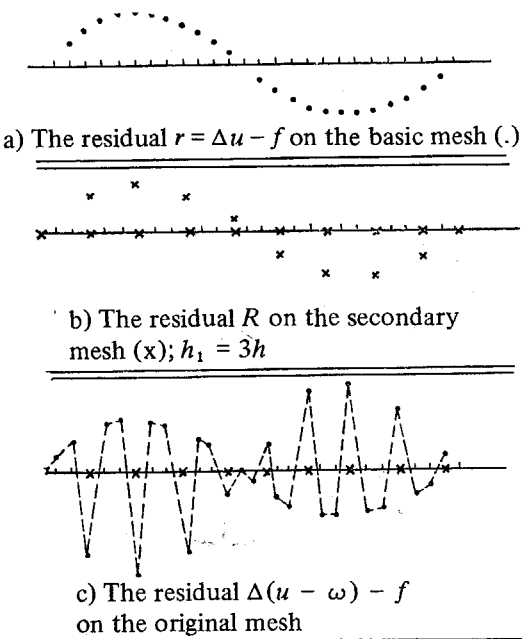
$$\|v^\nu\| \simeq \|v^0\| e^{-\varkappa\nu} \tag{9.4}$$

and $\varkappa$ is again independent of the number of mesh-points. Below we describe the algorithm of this method, its qualitative justification and examples illustrating its efficiency in practice. We also deduce the estimate (9.1) in the same way as in [18]. We limit ourselves to this weakest result, since its deduction is completely elementary. The more powerful proofs of [19] and [20] require special tools and estimates.

Qualitatively, the method is based on the following remark. Using, say, the simple iterations

$$u^{\nu+1} = u^\nu + \tau(\Delta u^\nu - f) \tag{9.5}$$

with $\tau \simeq 1/L$ it is easy to suppress the high-frequency component (corresponding to $\lambda \simeq L$) of the residual (3.3). As regards the low-frequency component ($\lambda \simeq l$), it decreases as slowly as $e^{-\nu l/L}$; and it is, in fact, the suppression of this component that represents the main difficulty of this problem. As shown for example by Fig. 5, this behaviour is not limited to the simple iterations (9.5). We also note that the eigenfunctions for $\lambda \simeq l$ are smooth with few changes of sign in the domain in question (the smoothness increases with decreasing λ); on the other hand, the eigenfunctions with $\lambda \sim L$ change rapidly. This qualitative feature is typical for the most general elliptic operators and their finite-dimensional approximations, and it is this feature that is utilized in the construction of the algorithm and determines the corresponding range of specific problems.



a) The residual $r = \Delta u - f$ on the basic mesh (.)

b) The residual R on the secondary mesh (x); $h_1 = 3h$

c) The residual $\Delta(u - \omega) - f$ on the original mesh

Fig. 8.

We now turn to the computational scheme of the algorithm. Let us start with an initial approximation u^0. We perform a small number of simple iterations (9.5), choosing τ to suppress the high-frequency component of the residual. As a result, the residual $r^\nu = \Delta u^\nu - f$ becomes a smooth function. This stage of the computation is qualitatively illustrated in Fig. 8a.

If it were possible to find a function w solving the equation $\Delta w = r^\nu$, the problem would be solved, for then $u = u^\nu - w$ gives $\Delta(u^\nu - w) = r^\nu + f - r^\nu = f$. At first sight the solution of the equation $\Delta w = r^\nu$ would appear to be no simpler than that of the problem $\Delta u = f$, but this is not so. The reason is that r^ν is smooth, and therefore so is w, Hence we can find w on a mesh whose step is greater than the original mesh-size, which is easier, as there are fewer vertices and the convergence of the iterations improves. Thus, we define the function R on the mesh with the greater step by setting $R = r^\nu$ on the common vertices. This is illustrated in Fig. 8b where the step of the auxiliary mesh is three times as large as the original step. We solve the difference equation

$$\Delta W = R, \tag{9.6}$$

where W and R are determined on the vertices of the supplementary mesh, Δ is the difference operator approximating the same differential operator[1] as Δ, but on the other mesh. From W we obtain the function w, which is defined on the original mesh, by some (say, linear) interpolation. We now amend the function u^ν: $\tilde{u}^\nu := u^\nu - w$.

We compute $\tilde{r}^\nu = \Delta\tilde{u}^\nu - f$; this function is shown in Fig. 8c. The following remark explains the picture: suppose that W_{3n} satisfies the difference equation on the secondary mesh

$$\frac{W_{3n-3} - 2W_{3n} + W_{3n+3}}{(3h)^2} = R_{3n} = r_{3n} \left(n = 1, 2, \ldots, \frac{N}{3} - 1\right),$$

and that w_n ($n = 0, 1, \ldots, N$) is the linear interpolation of W to the remaining vertices of the original mesh with the step h. Then

$$\frac{w_{n-1} - 2w_n + w_{n+1}}{h^2} = \begin{cases} 0 \text{ if } n \text{ is divisible by 3,} \\ 3R_n \text{ otherwise.} \end{cases}$$

From the point of view of the residual norm, $\tilde{u}^\nu$ is worse than u^ν (a rough estimate gives $\|\tilde{r}^\nu\| \simeq 1{\cdot}5\|r^\nu\|$). However, $\tilde{r}^\nu$ consists mainly of high-frequency eigenfunctions, and in the mean it is close to zero. Repeating the simple iterations (9.5) we suppress the high frequency part of $\tilde{r}^\nu$ etc. We remark that usually the step of the secondary mesh is only two or three times greater than the original step, and so the solution of (9.6) need not be a simple problem at all. Then we use similar method to determine W. This is, rather vaguely, the qualitative description of the method. The details are specified below, where we estimate its rate of convergence and describe the experience obtained in its practical application.

Estimate of the rate of convergence. We consider the application of our method to the solution of the difference equation

$$(\Delta u)_{n,m} = \left(\frac{\partial^2 u}{\partial x^2}\right)_{n,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)_{n,m} = f_{n,m}, \quad u|_\Gamma = \varphi \tag{9.7}$$

[1] The boundary conditions for W are the same as for u; they are, of course, homogeneous.

on a mesh of step $h = \pi/N$, where $N = 2^s$. To avoid confusion and facilitate the understanding of the following exposition, we fix our notation. We shall be concerned with two meshes: the basic one, whose step is h, and a secondary one of step $H = 2h$. All the objects defined on the basic mesh are marked by small letters and those on the secondary mesh are marked by capitals. When we pass from functions defined on one mesh to similar functions on the other we use the same letter of a different size. Thus, let $u^0_{n,m}$ be the initial approximation, whose residual r^0 is expanded in the Fourier series

$$r^0 = \sum_{p,\,q=1}^{N-1} c_{p,\,q}\varphi^{(p,\,q)}, \qquad \|r^0\| = \delta_0 = \Big[\sum_{p,\,q=1}^{N-1} c^2_{p,\,q}\Big]^{1/2}. \tag{9.8}$$

(Here $\varphi^{(p,\,q)}_{n,m} \simeq \sin\frac{pn\pi}{N}\sin\frac{qm\pi}{N}$; the exact form of $\varphi^{(p,\,q)}$ is used only once.)

We split the set of eigenfunctions into two subsets – the good, smooth functions (p and $q < N/2$) and the bad, rough functions (p or $q \geqslant N/2$). Accordingly, the space of all the mesh functions[1] splits into the orthogonal sum of two subspaces: a good one and a bad one. We sometimes mark the elements of these subspaces by the subscripts[2] X and Π, respectively. In this notation

$$r^0 = r_{\text{x}} + r_{\text{п}}, \quad \|r_{\text{x}}\| \leqslant \delta_0, \quad \|r_{\text{п}}\| \leqslant \delta_0, \quad \|r^0\|^2 = \|r_{\text{x}}\|^2 + \|r_{\text{п}}\|^2.$$

We perform ν iterations (9.5) with the fixed value of the parameter $\tau = \frac{1}{5}h^2 = \frac{1}{5}\frac{\pi^2}{N^2}$. In this case $|1 - \tau\lambda| < \rho = 0.6$ on the bad part of the spectrum and $|1 - \tau\lambda| < 1$ on the good part. Hence we obtain the functions

$$\begin{cases} r^\nu = \sum\limits_{p,\,q} c_{p,\,q}(1-\tau\lambda_{p,\,q})^\nu \varphi^{(p,\,q)}, \\ r^\nu = r^\nu_{\text{x}} + r^\nu_{\text{п}}, \quad \|r^\nu_{\text{x}}\| \leqslant \delta_0, \quad \|r^\nu_{\text{п}}\| \leqslant \rho^\nu \cdot \delta_0. \end{cases} \tag{9.9}$$

We consider now on the secondary mesh the function

$$R_{2i,\,2j} = r^\nu_{2i,\,2j}\left(i,\ j = 1,\ 2,\ \ldots,\ \frac{N}{2} - 1 = N' - 1\right) \tag{9.10}$$

and find the function $W_{2i,2j}$ satisfying the difference equation

$$(\triangle W)_{2i,\,2j} = R_{2i,\,2j} + \|R\|\mathscr{E}_{2i,\,2j}, \quad W|_\Gamma = 0 \tag{9.11}$$

Here E is an arbitrary function subject only to the condition $\|\mathscr{E}\| \leqslant \varepsilon$, where ε is to be determined later. Introducing $\mathscr{E}$ we emphasize that W is

[1] vanishing on the boundary,

[2] X and Π are the initial letters of the Russian words for good and bad. (Transl.)

found only approximately by some iterative process which diminishes ε^{-1} times the residual of the original approximation $W^0 = 0$. In accordance with the splitting $r^\nu = r^\nu_{\text{x}} + r^\nu_{\text{п}}$, $R = R^{\text{x}} + R^{\text{п}}$, we write W in the form $W = W^{\text{x}} + W^{\text{п}} + W^{\varepsilon}$ so that

$$\begin{gathered} \triangle W^{\text{x}} = R^{\text{x}}, \quad \triangle W^{\text{п}} = R^{\text{п}}, \\ \triangle W^{\varepsilon} = \|R\|\mathscr{E}, \quad W^{\text{x}} = W^{\text{п}} = W^{\varepsilon}|_\Gamma = 0. \end{gathered} \tag{9.12}$$

(The superscripts X and Π here indicate the origin of the function and not the subspace to which it belongs.)

The next lemmas are useful in what follows.

LEMMA 1. *Suppose that the function $z_{n,m}$ is defined on the vertices of the original mesh and that $Z_{2i,2j}$ is its 'projection' onto the secondary mesh (that is, $Z_{2i,2j} = z_{2i,2j}$). Then*

$$\|Z\| \leqslant 2\|z\|. \tag{9.13}$$

The proof is obvious:

$$\|Z\|^2 = H^2 \sum_{i,\,j=1}^{N'-1} Z^2_{2i,\,2j} = 4h^2 \sum_{i,\,j=1}^{N'-1} z^2_{2i,\,2j} \leqslant 4\|z\|^2.$$

LEMMA 2. *Let W be defined on the vertices of the secondary mesh and let w be the linear interpolation of W to the basic mesh. Then*

$$\|\Delta w\| \leqslant K\|\triangle W\|. \tag{9.14}$$

The constant K is independent of the mesh step and can easily be computed, but we are not interested in its precise value.

PROOF. We compute

$$(\Delta w)_{n,m} = \begin{cases} 2(\triangle W)_{n,\,m} & (n \text{ and } m \text{ even}) \\ \left(\frac{\partial^2 W}{\partial x^2}\right)_{n,\,m+1} + \left(\frac{\partial^2 W}{\partial x^2}\right)_{n,\,m-1} & (n \text{ even},\ m \text{ odd}) \\ \left(\frac{\partial^2 W}{\partial y^2}\right)_{n+1,\,m} + \left(\frac{\partial^2 W}{\partial y^2}\right)_{n-1,\,m} & (n \text{ odd},\ m \text{ even}) \\ 0 & (n \text{ and } m \text{ odd}). \end{cases} \tag{9.15}$$

It follows that $(\Delta w)_{n,m}$ is a combination of the values of $\triangle W$, $\frac{\partial^2 W}{\partial x^2}$ and $\frac{\partial^2 W}{\partial y^2}$ at the neighbouring vertices. Using the triangle inequality and the obvious estimates

$$\left\|\frac{\partial^2 W}{\partial x^2}\right\| \leqslant \|\triangle W\|, \qquad \left\|\frac{\partial^2 W}{\partial y^2}\right\| \leqslant \|\triangle W\|$$

(which follow from the fact that the operators $\frac{\partial^2}{\partial x^2}$ and $\frac{\partial^2}{\partial y^2}$ are negative-definite and commute) we can easily establish (9.14).

LEMMA 3. *Let W be defined on the secondary mesh and let w be its linear interpolation. Let $a_{p,q}$ $(p, q < N' = N/2)$ be the Fourier coefficients of Δw. Then*

$$a_{p,q} = A_{p,q} + \lambda'_p \frac{h^2}{4} A'_{p,q} + \lambda''_q \frac{h^2}{4} A''_{p,q}, \tag{9.16}$$

where $A_{p,q}$, $A'_{p,q}$ and $A''_{p,q}$ are, respectively, the Fourier coefficients of W, $\left(\frac{\partial^2 W}{\partial y^2}\right)$ and $\left(\frac{\partial^2 W}{\partial x^2}\right)$.

PROOF.

$$a_{p,q} = h^2 \sum_{n=1}^{N-1} \sum_{m=1}^{N-1} (\Delta w)_{n,m} \varphi_{n,m}^{(p,q)} =$$

$$= h^2 \Big\{ \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} (\Delta w \varphi)_{2i,2j} + \sum_{i=0}^{N'-1} \sum_{j=1}^{N'-1} (\Delta w \varphi)_{2i+1,2j} + \sum_{i=0}^{N'-1} \sum_{j=0}^{N'-1} (\Delta w \varphi)_{2i,2j+1} \Big\}.$$

Here, by virtue of (9.15), we omit the sum over 'odd-odd' vertices. Using (9.15) again, we have

$$a_{p,q} = 2h^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} (\Delta W)_{2i,2j} \varphi_{2i,2j} +$$

$$+ h^2 \sum_{i=0}^{N'-1} \sum_{j=1}^{N'-1} \varphi_{2i+1,2j} \left\{ \left(\frac{\partial^2 W}{\partial y^2}\right)_{2i,2j} + \left(\frac{\partial^2 W}{\partial y^2}\right)_{2i+2,2j} \right\} +$$

$$+ h^2 \sum_{i=1}^{N'-1} \sum_{j=0}^{N'-1} \varphi_{2i,2j+1} \left\{ \left(\frac{\partial^2 W}{\partial x^2}\right)_{2i,2j} + \left(\frac{\partial^2 W}{\partial x^2}\right)_{2i,2j+2} \right\} =$$

$$= 4h^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} (\Delta W)_{2i,2j} \varphi_{2i,2j} +$$

$$+ h^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} \left(\frac{\partial^2 W}{\partial y^2}\right)_{2i,2j} \{\varphi_{2i+1,2j} + \varphi_{2i-1,2j}\} - 2h^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} \left(\frac{\partial^2 W}{\partial y^2} \varphi\right)_{2i,2j} +$$

$$+ h^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} \left(\frac{\partial^2 W}{\partial x^2}\right)_{2i,2j} \{\varphi_{2i,2j+1} + \varphi_{2i,2j-1}\} - 2h^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} \left(\frac{\partial^2 W}{\partial x^2} \varphi\right)_{2i,2j} =$$

$$= H^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} (\Delta W)_{2i,2j} \varphi_{2i,2j} + h^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} \left(\frac{\partial^2 W}{\partial y^2}\right)_{2i,2j} h^2 \left(\frac{\partial^2 \varphi}{\partial x^2}\right)_{2i,2j} +$$

$$+ h^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} \left(\frac{\partial^2 W}{\partial x^2}\right)_{2i,2j} h^2 \left(\frac{\partial^2 \varphi}{\partial y^2}\right)_{2i,2j}.$$

Accordingly, we put

$$a_{p,q} = A_{p,q} + \frac{\lambda'_p h^2}{4} A'_{p,q} + \frac{\lambda''_q h^2}{4} A''_{p,q}, \tag{9.17}$$

where

$$\begin{cases} A_{p,q} = H^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} (\Delta W)_{2i,2j} \Phi_{2i,2j}^{(p,q)}, \\ A'_{p,q} = H^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} \left(\frac{\partial^2 W}{\partial y^2}\right)_{2i,2j} \Phi_{2i,2j}^{(p,q)}, \\ A''_{p,q} = H^2 \sum_{i=1}^{N'-1} \sum_{j=1}^{N'-1} \left(\frac{\partial^2 W}{\partial x^2}\right)_{2i,2j} \Phi_{2i,2j}^{(p,q)}. \end{cases} \tag{9.18}$$

We use here the fact that in the simple problem in question the values of the eigenfunctions of $\triangle$ and Δ on the coincident vertices of the two meshes are the same for p and $q \leqslant N' - 1$ (these 'good' p and q are the only ones for which (9.17) is used below). In effect, this is the only point in the proof where we make real use of the special form of the problem. The equality

$$\varphi_{2i,2j}^{(p,q)} = \Phi_{2i,2j}^{(p,q)} \tag{9.19}$$

is accidentally precise. But it is related to the completely non-accidental fact that, for small values of p and q, the functions φ and Φ approximate one and the same eigenfunction of the differential Laplace operator. The accuracy of the approximation increases with decreasing p and q. It is useful to note that the proof goes through with almost no changes if the 'good' part of the spectrum is determined by the condition $p, q < \xi N$, where $\xi > 0$ is an arbitrary small fixed number. This remark points to the reserves that can be used in generalizing the proof. The most difficult to generalize are the estimates below. Let us return to the function w, which by (9.12) can be represented in the form

$$w = w^{\mathrm{x}} + w^{\mathrm{n}} + w^{\mathrm{e}}.$$

Each of the terms on the right-hand side is the linear interpolation of W^{x}, W^{n} and W^{e}, respectively. By (9.10), (9.13) and (9.14), we have

$$\| \Delta w^{\mathrm{n}} \| \leqslant K \| \Delta W^{\mathrm{n}} \| = K \| R^{\mathrm{n}} \| \leqslant 2K \cdot \| r_{\mathrm{n}}^{\mathrm{v}} \| \leqslant 2K \rho^{\mathrm{v}} \delta_0,$$
$$\| w^{\mathrm{e}} \| \leqslant K \| \Delta W^{\mathrm{e}} \| \leqslant K \| R \|_{\varepsilon} \leqslant 2K \varepsilon \delta_0.$$

As to $\| \Delta w^{\mathrm{x}} \|$, we have to estimate it more accurately. First of all we decompose Δw^{x} in the orthogonal sum $\Delta w^{\mathrm{x}} = \Delta w_{\mathrm{x}}^{\mathrm{x}} + \Delta w_{\mathrm{n}}^{\mathrm{x}}$. For $\| \Delta w_{\mathrm{n}}^{\mathrm{x}} \|$ we use the rough estimate

$$\| \Delta w_{\mathrm{n}}^{\mathrm{x}} \| \leqslant \| \Delta w^{\mathrm{x}} \| \leqslant K \| \Delta W^{\mathrm{x}} \| = K \| R^{\mathrm{x}} \| \leqslant 2K \| r_{\mathrm{x}}^{\mathrm{v}} \| \leqslant 2K \delta_0. \tag{9.20}$$

To estimate $\Delta w_{\mathrm{x}}^{\mathrm{x}}$ we represent it as a sum of three terms according to

(9.17): $\Delta w_x^x = z + z' + z''$, where

$$z = \sum_{p,\,q=1}^{N'-1} A_{p,\,q}\varphi^{(p,\,q)}, \qquad z' = \sum_{p,\,q=1}^{N'-1} \frac{h^2}{4}\lambda'_p A'_{p,\,q}\varphi^{(pq)}, \qquad z'' = \sum_{p,\,q=1}^{N'-1} \frac{h^2}{4}\lambda''_q A''_{p,\,q}\varphi^{(p,\,q)}.$$

Note that

$$(\Delta W^x)_{2i,\,2j} = R^x_{2i,\,2j} = (r^\nu_x)_{2i,\,2j} = \sum_{p,\,q=1}^{N'-1} c_{p,\,q}(1-\tau\lambda_{p,\,q})^\nu \varphi^{(p,\,q)}_{2i,\,2j} = \sum_{p,\,q=1}^{N'-1} c_{p,\,q}(1-\tau\lambda_{p,\,q})^\nu \Phi^{(p,\,q)}_{2i,\,2j}.$$

From the orthonormality of the system $\Phi^{(p,\,q)}$ we conclude that

$$A_{p,\,q} = c_{p,\,q}(1-\tau\lambda_{pq})^\nu \quad (p,\,q = 1,\,2,\,\ldots,\,N'-1) \tag{9.21}$$

hence $z = r^\nu_x$, and we are done. Note that $A'_{p,q}$ and $A''_{p,q}$, are the Fourier coefficients in the expansions of $\frac{\partial^2 W}{\partial y^2}$ and $\frac{\partial^2 W}{\partial x^2}$, respectively. Therefore

$$\left\{\sum_{p,\,q=1}^{N'-1}(A'_{p,\,q})^2\right\}^{1/2} = \left\|\frac{\partial^2 W}{\partial y^2}\right\| \leqslant \|\Delta W^x\| = \|R^x\| \leqslant 2\|r^\nu_x\| \leqslant 2\delta_0, \tag{9.22}$$

and in exactly the same way

$$\left\{\sum_{p,\,q=1}^{N'-1}(A''_{p,\,q})^2\right\}^{1/2} = \left\|\frac{\partial^2 W}{\partial x^2}\right\| \leqslant \|\Delta W^x\| \leqslant 2\delta_0. \tag{9.23}$$

To sum up, we consider the residual of the corrected function $\tilde{u} = u^\nu - w$; we obtain

$$\tilde{r} = \Delta\tilde{u} - f = \Delta u^\nu - f - \Delta w = r^\nu - \Delta(w^x + w^\pi + w^\varepsilon) = r^\nu - (z + z' + z'') - \Delta w^\pi - \Delta w^\varepsilon - \Delta w^x_\pi = r^\nu_\pi - \{z' + z'' + \Delta w^\pi + \Delta w^\varepsilon + \Delta w^x_\pi\},$$

and we already have estimates for each of the terms in the last pair of braces. Now, starting with $\tilde{u}$ we perform ν additional simple iterations to obtain the function $u^{2\nu}$, whose residual is given by

$$r^{2\nu} = \tilde{r}^\nu_\pi - \{\tilde{z}' + \tilde{z}'' + \Delta\tilde{w}^\pi + \Delta\tilde{w}^\varepsilon + \Delta\tilde{w}^x_\pi\};$$

here every term in $r^{2\nu}$ represents a ν-times iterated component of $\tilde{r}$, for example, $\tilde{z}' = (E + \tau\Delta)^\nu z'$, etc. Let us estimate these terms.

1) $\|\Delta\tilde{w}^x\| \leqslant \rho^\nu \|\Delta w^x\| \leqslant 2K\delta_0\rho^\nu$ because this is a function in the bad subspace.

2) $\|\Delta\tilde{w}^\varepsilon\| \leqslant \|\Delta w^\varepsilon\| \leqslant 2K\varepsilon\delta_0$, because $\|E + \tau\Delta\| < 1$ and $\|\Delta w^\varepsilon\|$ satisfies (9.19).

3) $\|\Delta\tilde{w}^\pi\| \leqslant \|\Delta w^\pi\| \leqslant 2K\rho^\nu\delta_0$ for the same reasons.

4) $\tilde{z}' = (E + \tau\Delta)^\nu z' = \sum_{p,\,q=1}^{N'-1}(1-\tau\lambda_{p,\,q})^\nu \lambda'_p \frac{h^2}{4} A'_{p,\,q}\varphi^{(p,\,q)}$. We estimate the function $(1-\tau\lambda_{p,\,q})^\nu h^2\lambda_{p,\,q} = \left(1-\frac{h^2}{5}\lambda_{p,\,q}\right)^\nu h^2\lambda_{p,\,q}$ on the spectrum $0 \leqslant \lambda_{p,q} \leqslant 8/h^2$. By an elementary argument

$$\max_{0\leqslant\xi\leqslant 8}(1-0.2\xi)^\nu \xi = (1-0.2\xi_{\max})^\nu \xi_{\max} \leqslant \xi_{\max} = \frac{5}{\nu+1}.$$

Thus,

$$\|\tilde{z}'\| \leqslant \frac{5}{4}\frac{1}{\nu+1}\left\|\sum_{p,\,q=1}^{N'-1} A'_{p,\,q}\varphi^{(p,\,q)}\right\| \leqslant \frac{2.5}{\nu+1}\delta_0.$$

Similarly,

$$\|\tilde{z}''\| \leqslant \frac{2.5}{\nu+1}\delta_0.$$

Adding all the estimates we obtain

$$\|\Delta u^{2\nu} - f\| \leqslant \frac{5}{\nu+1}\delta_0 + 2K\delta_0\rho^\nu + 2K\varepsilon\delta_0 + 2K\delta_0\rho^\nu + \rho^{2\nu}\delta_0.$$

We choose $\nu(\varepsilon)$ so that $\frac{5}{\nu+1} + 4K\rho^\nu\rho^{2\nu} < K\varepsilon$. Then

$$\|\Delta u^{2\nu} - f\| \leqslant 3K\varepsilon\delta_0.$$

Starting with $u^{2\nu}$ we repeat all the computations described above to obtain the function $u^{4\nu}$ for which

$$\|\Delta u^{4\nu} - f\| \leqslant 3K\varepsilon\|\Delta u^{2\nu} - f\| \leqslant 9K^2\varepsilon^2\delta_0.$$

Now we choose ε such that $9K^2\varepsilon^2 \leqslant \varepsilon$, say, $\varepsilon = (9K^2)^{-1}$, and then we have

$$\|\Delta u^{4\nu} - f\| \leqslant \varepsilon\|\Delta u^0 - f\| = \varepsilon\delta_0.$$

The next lemma sums up the above argument; we emphasize that ε and $\nu(\varepsilon)$ are independent of the mesh size.

LEMMA. *To reduce the residual of the arbitrary initial approximation ε^{-1} times by the relaxation method it is sufficient*

a) *to perform $4\nu(\varepsilon)$ simple iterations,*

b) *twice interpolate W to the basic mesh, and*

c) *twice solve the finite difference Poisson equation on the secondary mesh with the accuracy ε. (By solving with accuracy ε we mean here the reduction of the residual of the original approximation ε^{-1} times* (see 9.11).)

Now let $Q(N,\varepsilon)$ denote the number of operations necessary for reducing the residual ε^{-1} times on the $(N \times N)$-mesh. The assertion of the lemma gives the inequality

$$Q(N,\varepsilon) \leqslant CN^2 + 2Q\left(\frac{N}{2},\varepsilon\right),$$

where C is independent of the mesh size (that is, of N). Expanding this estimate (which means that the problem on the $\frac{N}{2} \times \frac{N}{2}$ mesh is solved using the auxiliary $\frac{N}{4} \times \frac{N}{4}$ mesh etc.) we obtain

$$Q(N, \varepsilon) \leqslant CN^2 + 2C\frac{N^2}{4} + 4Q\left(\frac{N}{4}, \varepsilon\right) \leqslant \ldots$$
$$\ldots \leqslant CN^2\left(1 + \frac{1}{2} + \frac{1}{4} + \ldots + \right) \leqslant C_1 N^2.$$

Hence, at least for $N = 2^s$, we have proved the following theorem.

THEOREM. *To reduce by the relaxation method on the $N \times N$ mesh the residual of the initial approximation ε^{-1} times it is sufficient to perform at most C_1N^2 operations.*

The time for one iteration is $T = A \cdot N^2$, and so the assertion of the theorem is equivalent to the claim that in the relaxation method the residual decreases on average according to the formula

$$\| r^\nu \| \simeq \| r^0 \| e^{-\varkappa\nu}$$

where $\varkappa$ is independent of the number of mesh points.

A practical form of the method. The estimate of the rate of convergence obtained above is very rough. It would not be wise, therefore, to organize the computation strictly on the lines of the proof (that is, for example, to use the estimates obtained for ν, ε etc.) The form of the method used in practical computations is similar, to, but not identical with that used in deriving the estimate. This practical form is based on experience gained in solving a number of problems. We list the relevant recommendations below.

1. For the basic iteration process we take Seidel's method (see § 3, (3.20)) rather than the simple iteration (9.5), because it suppresses the high-frequency component of the error more efficiently. Other iterations are applicable, in particular, those using the sweep method. We mention here a situation where the choice of the basic iteration process has to be more specific: suppose that we have to solve the problem

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial y^2} = f$$

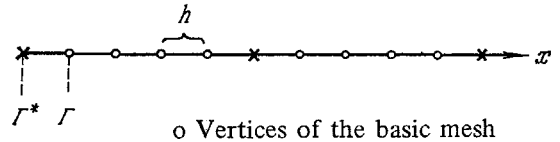
with $a \ll b$ (more precisely, if the mesh steps in x and y are not the same, with $a/\Delta x^2 \ll b/\Delta y^2$). In this case the simple, or Seidel's iterations, liquidate well the harmonics that oscillate rapidly in y, but the damping of components that are not smooth in x requires considerably more iterations than in the case $a/\Delta x^2 \approx b/\Delta y^2$.

2. The step of the secondary mesh is usually more than twice the size of the original step (the computations were performed with $H = 3h$, $4h$ or $5h$). In current practice ($N \sim 10^2$) it is sufficient to use two auxiliary meshes.

3. The number ν of iterations on the basic mesh after which we revert to the auxiliary mesh is usually taken within the limits $\nu \simeq 1.5 \div 3\ H/h$. This choice is explained by the nature of the residuals of the corrected function (see Fig. 8c): we require enough time for the residual at the point of the auxiliary mesh to be averaged out by the residuals at the neighbouring points (any iteration process for elliptic equations can be interpreted as a process of averaging residuals).

4. Computing the residuals on the basic mesh (with the problem posed on the first auxiliary mesh) gives sufficient control over the achieved accuracy.

We now describe the results of certain computations and further specify the details of the algorithm. In [17] there is a description of the experience with the solution of the problem



o Vertices of the basic mesh

x Vertices of the auxiliary mesh ($h^* = 5h$)

The boundary condition

$v + h\frac{\partial u}{\partial x} = 0|_{\Gamma^*}$, that is, $v = 0|_\Gamma$

Fig. 9

$\Delta u = f$ on the (40 × 48)-mesh (with equal steps in the directions of x and y). The basic iteration method was taken to be Seidel's, and convergence was accelerated by the use of one supplementary mesh with $H = 5h$. Not all the vertices of the supplementary mesh coincided with the vertices of the basic mesh; moreover, the supplementary mesh covered a wider area than the basic mesh. In this case we used boundary conditions of the third kind on the supplementary mesh so that the interpolated function satisfied the homogeneous boundary conditions of the original problem, as explained in Fig. 9. The number ν of iterations after which we reverted to the auxiliary mesh was taken within the limits 7 to 12. In all these cases the mean convergence (taking into account the time lost for reverting to the auxiliary mesh) corresponded to

$$\| r^t \| \simeq \| r^0 \| e^{-0.185 \cdot t/T},$$

where T is the time for one Seidel iteration (in the case in question it is basically given by[1] four additions and one division by 4). With this form of the relaxation method the residual of the initial approximation was

[1] About 3–4 iterations per second (on *M*-20).

reduced 10^3 times in less than 10 seconds (these computations were performed in 1960 in connection with a problem of the dynamics of an incompressible gas). The results were the same for the first and second boundary-value problems.

Table 5 illustrates the convergence of iterations in the solution of the difference equation

$$u_{xx} + 2bu_{xy} + u_{yy} = 0, \quad u|_\Gamma = \varphi,$$

b \ t	0	8 T	16 T	24 T	32 T	40 T	$\varkappa$
0	36.3	1.18	.048	.0029	—	—	0.39
0.5	36.3	1.48	.063	.0030	.00016	—	0.38
0.75	36.3	1.88	.091	.0053	.00032	.000029	0.36
0.875	36.3	2.24	.131	.012	—	—	0.32
0.9375	36.3	2.59	.20	.026	—	—	0.30

Table 5

on the (54 × 54) mesh.[1] The first supplementary mesh had 18 × 18 points, the second had 6 × 6. The scheme of the basic iteration cycle was as follows:

I. Six Seidel iterations are performed on the basic mesh

$$u^{\nu+1}_{n,m} = \frac{1}{4-2b}\{(1-b)(u^{\nu+1}_{n-1,m} + u^{\nu+1}_{n,m-1} + u^{\nu}_{n,m+1} + u^{\nu}_{n+1,m}) + b(u^{\nu+1}_{n-1,m-1} + u^{\nu}_{n+1,m+1}).$$

II. The residual R at the vertices of the supplementary 18 × 18 mesh ($H = 3h$) is computed and, starting with $W = 0$, six of the same iterations are performed for the equation

$$W_{xx} + 2bW_{xy} + W_{yy} = R, \quad W|_\Gamma = 0.$$

III. The residual $\widetilde{R}$ at the vertices of the (6 × 6)-mesh ($\widetilde{H} = 9h$) is computed and twenty iterations are used to solve the equation

$$\widetilde{W}_{xx} + 2b\widetilde{W}_{xy} + \widetilde{W}_{yy} = \widetilde{R}, \quad \widetilde{W}|_\Gamma = 0.$$

IV. The function $W := W - \widetilde{W}_{\text{interpolated}}$ is corrected and four more iterations on the 18 × 18 mesh are performed.

V. The function $u := u - W_{\text{interpolated}}$ is corrected.

The time for this cycle is $\simeq 8T$, where T is the time for one Seidel iteration (in this case two multiplications and five additions per point).

The table shows the decrease of $\|r^\nu\|$ for various values of b and the average value of $\varkappa(b)$ in the formula

$$\|r^\nu\| \simeq \|r_0\| e^{-\nu\varkappa}$$

(in all the computations the initial approximation u^0 was zero inside the

[1] The computations were performed in 1965 and reported by the author at the second All-Union Conference on Applied Mathematics.

domain and equal to the prescribed values on the boundary). We compare these computations with the method of alternating directions in its most efficient form, using the exact theory of the choice of the iteration parameters. A cycle of 8 double iterations in this case ($\eta \simeq 0.001$) reduces the residual of the initial approximation by 3.5 × 10^3 times (a cycle of 16 iterations – by 4 × 10^7 times). From table 5 it is clear that a similar result is obtained by the relaxation method in 24 iterations. However, an objective comparison must take into account the time needed for one iteration. For the method of alternating directions this is given by 16 additions and 16 multiplications per point. Taking into account the time for these operations (on the *M*-20; the result for other computers would be similar), the time-cost of one iteration of the alternating direction method is about six times as high as for one Seidel iteration. Hence the relaxation method is in this case at least twice as efficient as the method of alternating directions. This, however, is only true in the simplest case of the equation $\Delta u = f$ with equal steps in x and y. For the equation

$$\frac{\partial}{\partial x} a(x, y) \frac{\partial u}{\partial x} + \frac{\partial}{\partial y} c(x, y) \frac{\partial u}{\partial y} = f$$

with complicated functions a and b the duration of the iteration is basically determined by the number of times we refer to the functions $a(x, y)$ and $c(x, y)$ (if the volume of the operative memory does not permit the storage of their values) and the ratio of the time-costs will be 2 rather than 6. Nevertheless, if we wish to stay within the framework of commuting operators, for which the exact theory of the alternating method of directions holds, we may only consider the coefficients $a(x)$ and $c(y)$; these are, as a rule, easily stored, and the ratio of the time-costs increases again.[1]

As for a name of this method, it is related to the well-known relaxation method of Southwell, which has been used efficiently in the past in hand calculations by groups of experienced computers. Its idea is to correct the solution close to a high value of residual; the procedure depends on the experience of the computer and leads to the rapid reduction of residuals. The rules for the choice of the place and size of the correction were later formalized, but they were not used in machine computations because the expenditure of machine time on their complicated logic rendered the whole method not very efficient.

It is natural to assume that the basis of the computer's experience is the art of guessing a rough solution to the Poisson equation with the residual on the right-hand side. The same idea lies at the basis of the method described.

[1] It is worth mentioning that the calculation of the time needed assumes here a qualified assembly programmer. Using languages such as ALGOL or FORTRAN in combination with a mediocre compiler we end up with the ratio that is closer to 2 because of the extravagant expense of time on the logic. (This time would be negligible in the case of the assembly programmer, even is he is not very experienced.)

§ 10. The method of minimal residuals

The main idea of the method has been described in § 3.11. The solution of a self-adjoint difference equation

$$(Du)_{n,m} = f_{n,m}$$

minimizes the quadratic form

$$\sum_n \sum_m (Du)_{n,m} u_{n,m} - 2 \sum_n \sum_m u_{n,m} f_{n,m} \tag{10.1}$$

(the sum is taken over the interior mesh points). The iterations of the method of minimal residuals proceed by the following scheme. Starting with an approximation u^ν, we compute in some way a function v^ν, satisfying the homogeneous boundary conditions of the problem. The next approximation is taken in the form $u^{\nu+1} = u^\nu + \tau v^\nu$, where τ is chosen to minimize the quadratic form (10.1). Various forms of this method correspond to the various ways of obtaining direction v of the descent (this has a decisive influence on the efficiency). The simplest way

$$v^\nu = Du^\nu - f \quad (v = r^\nu) \tag{10.2}$$

leads to very slow convergence, as in the method of simple iterations. Therefore in this form the method is not of great interest. A more general form is usually considered, where the direction of the descent is given by the equation

$$Bv^\nu = Du^\nu - f. \tag{10.3}$$

The operator B must satisfy the following two conditions:

1) it must be easily invertible, that is, the solution of (10.3) for v must be substantially easier than the original problem;

2) it must, as far as possible, accelerate the rate of convergence of the iteration process. Writing the iteration step in the form

$$u^{\nu+1} = u^\nu + \tau B^{-1}(Du^\nu - f), \tag{10.4}$$

we observe the obvious similarity with the above arguments. The only difference is that the choice of τ is based not on an estimate of the spectrum of $D^{1/2}B^{-1}D^{1/2}$, but on the objective test of minimizing the quadratic form (10.1). In the first case the step is independent of the current approximation u^ν, while in the second case it depends on it in an essential way. This is reflected in the special feature of the method of minimal residuals: its actual convergence turns out to be much better than predicted by the theory. The estimate of the convergence is in terms of the spectral bounds $0 < l < L$ of the operator $D^{1/2}B^{-1}D^{1/2}$. The convergence is not slower than

$$\left(\frac{L-l}{L+l}\right)^\nu \simeq \exp\left(-\nu \frac{2l}{L}\right). \tag{10.5}$$

Although this is an upper estimate, it is not too coarse: one can work with an initial approximation u^0 (which differs from the solution by the sum of two eigenfunctions corresponding to the end-points of the spectrum) such that the rate of convergence is given from the beginning by (10.5). In practice, the rate for the first iterations is usually much higher than (10.5). Later it slows down, but by the time it reaches (10.5) the desired accuracy of the approximation u^ν has usually already been achieved.

As for the operator B, we have available at present several concrete constructions giving quite efficient results. We note that the ideal choice would be $B = D$, in which case the process would be complete in a single iteration. Although this is of no practical value, it indicates that B should be as close to D as possible. Let us consider the constructions for B that are available at present (these constructions have already been considered in § 8).

1. $B \equiv \Delta$. This is the best choice of B from the point of view of the second requirement, and the worst for an easy inversion. It may be recommended for solving the boundary-value problems in a rectangle. The equation $\Delta v = Du - f$ can be solved by the method of alternating directions with a small number of iterations, using the exact theory for the choice of parameters. We remark that the boundary conditions for v are here (as well as in the remaining cases) the homogeneous boundary conditions of the original problem.

2. In [21] Godunov and Prokopov, starting with $B \equiv \Delta$, came to the conclusion that rather than using a large number of iterations to obtain a sufficiently accurate solution of $\Delta v = Du - f$, it is better to use one iteration. In effect, this is equivalent to the construction

$$Bv \equiv \left(E - \sigma \frac{\partial^2}{\partial x^2}\right)\left(E - \sigma \frac{\partial^2}{\partial y^2}\right) v. \tag{10.6}$$

The parameter σ is determined by the condition that l/L should be as close to 1 as possible (we recall that l and L are the spectral bounds of $D^{1/2}B^{-1}D^{1/2}$). The latter problem is usually not very accurately solved, but the numerical experiments in [21] show that relatively large variations of σ have practically no effect on the efficiency of the whole process. In the same paper it is suggested (and confirmed by computations) that the efficiency can be improved by varying σ with the iterations. In practice, the efficiency of this method is comparable to that of the method of alternating directions with the optimal choice of the iteration parameters, although the theoretical rate of convergence is the same as for Richardson's method.

3. The construction

$$B \equiv \left\{E + \sigma\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)\right\}\left\{E - \sigma\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)\right\} \tag{10.7}$$

has been proposed by Marchuk and is subject to a theoretical and practical

scrutiny in [10]. The results here are completely analogous (both from the theoretical and practical aspect) to the results for the construction of B (10.6). The parameter σ is chosen in [10] experimentally; it can be varied in a fairly wide band of values without a marked effect on the efficiency.

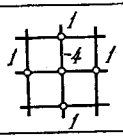
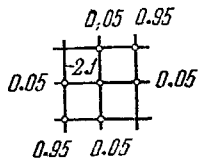
As for the corresponding class of concrete problems, we may remark that (10.6) may be used for solving self-adjoint equations (on condition that the finite-difference equations are also self-adjoint) in fairly general domains for the first boundary-value problem, and in the domains bounded by segments parallel to the coordinate axes for the third boundary-value problem.

Appendix

The author has recently performed some experiments aimed at clarifying the real efficiency of the relaxation method of § 9. They consisted in solving the first boundary-value problem in the $(\pi \times \pi)$-square for various equations. The right-hand side $f(x, y)$ stayed the same in all examples, as did the boundary conditions (1 on one of the sides of the square, 0 on the others). The size of the basic mesh was $h_0 = \pi/108$, and the steps of the first and second supplementary meshes were, respectively, $h_1 = \pi/36$ and $h_2 = \pi/12$.

For the basic iteration process we took the following procedure: in determining $u^{\nu+1}_{n,m}(m = 1, \ldots, N-1)$ all the quantities $u_{n,m}$ and $u_{n-1,m}$ in the difference equation were marked by the iteration superscript $\nu + 1$, and all the $u_{n+1,m}$ entries were marked by the superscript ν. Thus, $u^{\nu+1}_{n,m}(m = 1, \ldots, N-1)$ are determined by simultaneous sweeps in m; there are no specific iteration parameters. The results of the experiments are given in Table 6 which shows the equation, the stencil of the difference scheme, and the number n_1 of the basic iterations after which the process reverted to the supplementary mesh:

Table 6

No.	Equation	Stencil	n_1	$\|r^\nu\|$
1 2 3	$u_{xx} + u_{yy} = f$		4 8 10	$\sim e^{-0.38\nu}$ $e^{-0.42\nu}$ $e^{-0.18\nu}$
4	$u_{xx} + 1.9u_{xy} + u_{yy} = f$		10	$\sim e^{-0.38}$

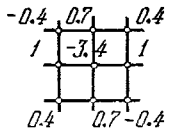
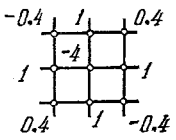
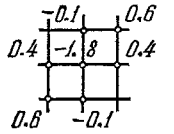
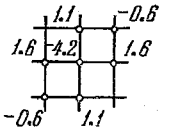
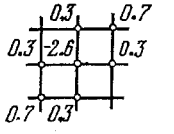
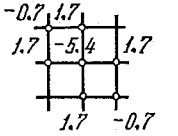
No.	Equation	Stencil	n_1	$\|r^\nu\|$
5	$u_{xx} + 1.6u_{xy} + 0.7u_{yy} = f$		10	$e^{-0.12\nu}$
6	$u_{xx} + 1.6u_{xy} + u_{yy} = f$		10	$e^{-0.18\nu}$
7 8	$u_{xx} + 1.2u_{xy} + 0.5u_{yy} = f$		8 10	$e^{-0.13\nu}$ $e^{-0.125\nu}$
9	$u_{xx} - 1.2u_{xy} + 0.5u_{yy} = f$		10	$e^{-0.13\nu}$
10	$u_{xx} + 1.4u_{xy} + 0.5u_{yy} = f$		8	$e^{-0.37\nu}$
11	$u_{xx} + 1.4u_{xy} + u_{yy} = f$		8	$e^{-0.18\nu}$

Table 7

No.	Example 3	Example 6	Example 10
0	355.	355.	355.
1	19.5	19.6	21.4
2	.061	.33	.24
3	.00058	.044	.0043
4	.0000064	.0085	.00012
5	.000000080	.0018	.0000043
6		.00042	.00000016
7		.00010	.0000000051
8		.000024	

The average efficiency per iteration is given by the expression

$\| r^{\nu} \| \simeq e^{-\gamma \cdot \nu}$; where γ is computed by the formula $\gamma = \frac{1}{\nu} \ln \|r^*\|$ and

$\|r^0\|$ is the initial residual, $\|r^*\|$ the residual of the final approximation, and ν the effective number of iterations (counting all the operations relating to the use of the auxiliary meshes as 2.5 iterations on the basic mesh per cycle).

Table 7 gives an idea of the nature of decrease of residuals with iterations in Examples 3, 6 and 10; here 0 indicates the residual of the initial approximation and 1 the residual of the function obtained after n_1 iterations on the basic mesh. This is followed by a correction, using auxiliary meshes and n_1 more iterations on the basic mesh; the residual of the resulting function u is marked 2, etc.

References

[1] V. I. Lebedev and V. N. Finogenov, Ordering the iteration parameters in the cyclic Chebyshev iterative method, Zh. Vychisl. Mat. i Mat. Fiz., **11** (1971), 425–438. MR **44** # 3479.
= USSR Computational Math. and Math. Phys. **11**:2 (1971), 155–170.

[2] V. I. Lebedev, Iterative methods of solving operator equations with spectrum located on several segments, A paper of the same author and title appears Zh. Vychisl. Mat. i Mat. Fiz. **9** (1969), 1247–1252. MR **42** # 7052.
= USSR Computation Math. and Math. Phys. **9**:6 (1969), 17–24.

[3] D. Young, Iterative methods for solving partial difference equations of elliptic type, Trans. Amer. Math. Soc. **76** (1954), 92–111.

[4] V. A. Magarik, A relaxation factor for systems with Young matrices, Zh. Vychisl. Mat. i Mat. Fiz **6** (1966), 824–830. MR **34** # 3768.
= USSR Computational Math. and Math. Phys. **6**: 5 (1966), 51–59.

[5] D. W. Peaceman and H. H. Rachford, The numerical solution of parabolic and elliptic differential equations, SIAM J. Appl. Math. **3** (1955), 28–42.

[6] J. Douglas and H. H. Rachford, On the numerical solution of heat conduction problems in two and three space variables, Trans. Amer. Math. Soc. **82** (1956), 421–439.

[7] S. K. Godunov and V. S. Ryaben'kii, *Vvedenie v teoriyu raznostnykh skhem*, Izd. Fiz. -Mat. Lit., Moscow 1962. MR **29** # 724.
Translation: Theory of difference schemes. An introduction. North-Holland, Amsterdam; Interscience, New York 1964. MR **31** # 5346.

[8] E. L. Wachspress, Extended application of alternating direction implicate iteration model problem theory, SIAM J. Appl. Math. **11** (1963), 994–1016.

[9] G. Birkhoff, R. S. Varga and D. Young, Alternating direction implicate methods, Advances in Computers **3**, 189–265, Academic Press, New York–London 1962.

[10] V. P. Il'in, On certain numerical experiments concerning iterative methods for difference Laplace equations, *Chislennye metody mekhaniky sploshnoi sredy* (Numerical methods of continuum mechanics, Novosibirsk), **1** (1970), 31–51.

[11] A. A. Samarskii, *Vvedenie v teoriyu raznostnykh skhem* (Introduction to the theory of difference schemes), Nauka, Moscow 1971.

[12] O. B. Widlund, On the effects of scaling of the Peaceman–Rachford method, Conf. on Numer. Solution of Differential Equations, Springer–Verlag, Berlin-Heidelberg-New York 1969. MR **44** # 1245.

[13] E. G. D'yakonov, The construction of iterative methods using spectrally equivalent operators, Zh. Vychisl. Mat. i Mat. Fiz., **6**: 1 (1966), 12–34. MR **33** # 6807.
= USSR Computational Math. and Math. Phys. **6** (1966), 14–46.

[14] A. N. Konovalov, Numerical solution of a mixed problem in elasticity theory, Zh. Vychisl. Mat. i Mat. Fiz., **9** (1969), 469–473. MR **40** # 6838.
= USSR Computational Math. and Math. Phys. **9**: 2 (1969), 296–304.

[15] R. W. Hockney, The potential calculation and some applications, Methods Comp. Phys. **9** (1970), 135–211.

[16] J. E. Gunn, The numerical solution of $\nabla a \nabla u = f$ by a semi-explicite alternating direction technique, Numer. Math. **6** (1964), 243–249.

[17] R. P. Fedorenko, A relaxation method for solving elliptic difference equations, Zh. Vychisl. Mat. i Mat. Fiz., **1** (1961), 922–927. MR **25** # 766.
= USSR Computational math. and Math. Phys. 1962, No. 4, 1092–1096.

[18] R. P. Fedorenko, The rate of convergence of an iterative process, Zh. Vychisl. Mat. i Mat. Fiz., **4** (1964), 559–564. MR **31** # 6386.
= USSR Computational Math. and Math. Phys. **4**: 3 (1964), 227–235.

[19] N. S. Bakhvalov, On the convergence of a relaxation method with natural constraints on the elliptic operator, Zh. Vychisl. Mat. i Mat. Fiz., **6** (1966), 861–885.
= USSR Computational Math. and Math. Phys. **6**: 5 (1966), 101–135.

[20] G. P. Astrakhantsev, An iterative method for solving elliptic difference problems, Zh. Vychisl. Mat. i Mat. Fiz., **11** (1971), 439–448. MR **44** # 1239.
= USSR Computational Math. and Math. Phys. **11**: 2 (1971), 171–182.

[21] S. K. Godunov and G. P. Prokopov, On the solution of the difference Laplace equation, Zh. Vychisl. Mat. i Mat. Fiz., **9** (1969), 462–468. MR **41** # 4832.
= USSR Computational Math. and Math. Phys. **9**: 2 (1969), 285–292.

Received by the Editors,
14 December 1972

Translated by P. Stefan.